An Efficient Solution to the Five-Point Relative Pose Problem

David Nistér
Sarnoff Corporation
CN5300, Princeton, NJ 08530
dnister@sarnoff.com

Abstract

An efficient algorithmic solution to the classical five-point relative pose problem is presented. The problem is to find the possible solutions for relative camera motion between two calibrated views given five corresponding points. The algorithm consists of computing the coefficients of a tenth degree polynomial and subsequently finding its roots. It is the first algorithm well suited for numerical implementation that also corresponds to the inherent complexity of the problem. The algorithm is used in a robust hypothesise-and-test framework to estimate structure and motion in real-time.

1. Introduction

Reconstruction of camera positions and scene structure based on images of scene features from multiple viewpoints has been studied for over two centuries, first by the photogrammetry community and more recently in computer vision. In the classical setting, the intrinsic parameters of the camera, such as focal length, are assumed known a priori. This calibrated setting is where the five-point problem arises. Given the images of five unknown scene points from two distinct unknown viewpoints, what are the possible solutions for the configuration of the points and cameras? Clearly, only the relative positions of the points and cameras can be recovered. Moreover, the overall scale of the configuration can never be recovered solely from images. Apart from this ambiguity, the five-point problem was proven by Kruppa [14] to have at most eleven solutions. This was later improved upon [2, 3, 5, 16, 11], showing that there are at most ten solutions and that there are ten solutions in general (including complex ones). The ten solutions correspond to the roots of a tenth degree polynomial. However, Kruppa's method requires the non-trivial operation of finding all intersections between two sextic curves and there is no previously known practical method of deriving the coefficients of the tenth degree polynomial in the general case. A few algorithms suitable for numerical implementation have also been devised. In [28] a 60×60 sparse matrix is built, which is subsequently reduced using linear algebra to a 20×20 non-symmetric matrix whose eigenvalues and eigenvectors encode the solution to the problem. In [21] an efficient derivation is given that leads to a thirteenth degree polynomial whose roots include the solutions to the five-point problem. The solution presented in this paper is a refinement of this. A better elimination that leads directly in closed form to the tenth degree polynomial is used. Thus, an efficient algorithm that corresponds exactly to the intrinsic degree of difficulty of the problem is obtained.

For the structure and motion estimation to be robust and accurate in practice, more than five points have to be used. The classical way of making use of many points is to minimise a least squares measure over all points, see for example [13]. Our intended application for the five-point algorithm is as a hypothesis generator within a random sample consensus scheme (RANSAC) [6]. Many random samples containing five point correspondences are taken. Each sample yields a number of hypotheses for the relative orientation that are scored by a robust statistical measure over all points in two or more views. The best hypothesis is then refined iteratively. Such a hypothesise-and-test architecture has become the standard way of dealing with mismatched point correspondences [26, 31, 10, 18] and has made automatic reconstructions spanning hundreds of views possible [1, 22, 19].

The requirement of prior intrinsic calibration was relaxed in the last decade [4, 9, 10], leading to higher flexibility and less complicated algorithms. So, why consider the calibrated setting? Apart from the theoretical interest, one answer to this question concerns stability and uniqueness of solutions. Enforcing the intrinsic calibration constraints often gives a crucial improvement of both the accuracy and robustness of the structure and motion estimates. Currently, the standard way of achieving this is through an initial uncalibrated estimate followed by iterative refinement to bring the estimate into agreement with the calibration constraints. When the intrinsic parameters *are* known a priori, the five-point algorithm is a more direct way of enforcing the calibration constraints exactly and obtaining a Euclidean reconstruction. The accuracy and robustness improvements gained by enforcing the calibration constraints are particu-

Prepared through collaborative participation in the Robotics Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0012. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

larly significant for planar or near planar scenes and scenes that *appear* planar in the imagery. The uncalibrated methods fail when faced with coplanar scene points, since there is then a continuum of possible solutions. It has been proposed to deal with this degeneracy using model selection [27, 23], switching between a homographic model and the general uncalibrated model as appropriate. In the calibrated setting, coplanar scene points only cause at most a two-fold ambiguity [15, 17]. With a third view, the ambiguity is in general resolved. In light of this, a RANSAC scheme that uses the five-point algorithm over three or more views is proposed. It applies to general structure but also continues to operate correctly *despite* scene planarity, without relying on or explicitly detecting the degeneracy. In essence, the calibrated model can cover both the planar and general structure cases seamlessly. This gives some hope of dealing with the approximately planar cases, where neither the planar nor the uncalibrated general structure model applies well.

The rest of the paper is organised as follows. Section 2 establishes some notation and describes the constraints used in the calibrated case. Section 3 presents the five-point algorithm. Section 4 discusses planar degeneracy. Section 5 outlines the RANSAC schemes for two and three views. Section 6 gives results and Section 7 concludes.

2. Preliminaries and Notation

Image points are represented by homogeneous 3-vectors q and q' in the first and second view, respectively. World points are represented by homogeneous 4-vectors Q. A perspective view is represented by a 3×4 camera matrix P indicating the image projection $q \sim PQ$, where $\sim$ denotes equality up to scale. A view with a finite projection centre can be factored into $P = K\,[R \mid t]$, where K is a 3×3 upper triangular calibration matrix holding the intrinsic parameters and R is a rotation matrix. Let the camera matrices for the two views be $K_1\,[I \mid 0]$ and $P = K_2\,[R \mid t]$. Let $[t]_\times$ denote the skew symmetric matrix

$$[t]_\times = \begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{bmatrix} \tag{1}$$

so that $[t]_\times\, x = t \times x$ for all x. Then the fundamental matrix is

$$F \equiv K_2^{-\top}\,[t]_\times\, R K_1^{-1}. \tag{2}$$

The fundamental matrix encodes the well known coplanarity, or epipolar constraint

$$q'^{\top} F q = 0. \tag{3}$$

If K_1 and K_2 are known, the cameras are said to be calibrated. In this case, we can always assume that the image points q and q' have been premultiplied by K_1^{-1} and K_2^{-1}, respectively, so that the epipolar constraint simplifies to

$$q'^{\top} E q = 0, \tag{4}$$

where the matrix $E \equiv [t]_\times\, R$ is called the essential matrix. Any rank-2 matrix is a possible fundamental matrix. An essential matrix has the additional property that the two non-zero singular values are equal. This leads to the following important cubic constraints on the essential matrix, adapted from [25, 5, 16, 21]:

Theorem 1 *A real non-zero 3×3 matrix E is an essential matrix if and only if it satisfies the equation*

$$EE^{\top}E - \frac{1}{2}\mathit{trace}(EE^{\top})E = 0. \tag{5}$$

This property will help us recover the essential matrix. Once the essential matrix is known, R, t and the camera matrices can be recovered from it.

3. The Five-Point Algorithm

In this section the five-point algorithm is described, first in a straightforward manner. Recommendations for an efficient implementation are then given in Section 3.2. Each of the five point correspondences gives rise to a constraint of the form (4). This constraint can also be written as

$$\tilde{q}^{\top}\tilde{E} = 0, \tag{6}$$

where

$$\tilde{q} \equiv \begin{bmatrix} q_1q'_1 & q_2q'_1 & q_3q'_1 & q_1q'_2 & q_2q'_2 & q_3q'_2 & q_1q'_3 & q_2q'_3 & q_3q'_3 \end{bmatrix}^{\top} \tag{7}$$

$$\tilde{E} \equiv \begin{bmatrix} E_{11} & E_{12} & E_{13} & E_{21} & E_{22} & E_{23} & E_{31} & E_{32} & E_{33} \end{bmatrix}^{\top} \tag{8}$$

By stacking the vectors $\tilde{q}^{\top}$ for all five points, a 5×9 matrix is obtained. Four vectors $\tilde{X}, \tilde{Y}, \tilde{Z}, \tilde{W}$ that span the right nullspace of this matrix are now computed. The most common way to achieve this is by singular value decomposition [24], but QR-factorisation as described in Section 3.2 is much more efficient. The four vectors correspond directly to four 3×3 matrices X, Y, Z, W and the essential matrix must be of the form

$$E = xX + yY + zZ + wW \tag{9}$$

for some scalars x, y, z, w. The four scalars are only defined up to a common scale factor and it is therefore assumed that $w = 1$. Note here that the algorithm can be extended to using more than 5 points in much the same way as the uncalibrated 7 and 8-point methods. In the overdetermined case, the four singular vectors X, Y, Z, W that correspond to the four smallest singular values are used. By inserting (9) into the nine cubic constraints (5) and performing Gauss-Jordan elimination with partial pivoting we obtain the equation system

A	x^3	y^3	x^2y	xy^2	x^2z	y^2z	x^2	y^2	xyz	xy	xz^2	xz	x	yz^2	yz	y	1
(a)	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	[3]
(b)		1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	[3]
(c)			1							.	.	.	.	.	.	.	[3]
(d)				1						.	.	.	.	.	.	.	[3]
(e)					1					.	.	.	.	.	.	.	[3]
(f)						1				.	.	.	.	.	.	.	[3]
(g)							1			L	.	.	.	M	N	O	[3]
(h)								1		P	Q	R	S	.	.	.	[3]
(i)									1	.	.	.	.	.	.	.	[3]

where . and $L, \ldots, S$ denote some scalar values and $[n]$ denotes a polynomial of degree n in the variable z. Note that the elimination can optionally be stopped two rows early. Further, define the additional equations

$$(j) \equiv (e) - z(g) \quad (10)$$
$$(k) \equiv (f) - z(h) \quad (11)$$
$$(l) \equiv (d) - x(h) + P(c) + zQ(e) + R(e) + S(g) \quad (12)$$
$$(m) \equiv (c) - y(g) + L(d) + zM(f) + N(f) + O(h). \quad (13)$$

We now have the five equations

$$(i) = xy[1] + x[2] + y[2] + [3] = 0 \quad (14)$$
$$(j) = xy[1] + x[3] + y[3] + [4] = 0 \quad (15)$$
$$(k) = xy[1] + x[3] + y[3] + [4] = 0 \quad (16)$$
$$(l) = xy[2] + x[3] + y[3] + [4] = 0 \quad (17)$$
$$(m) = xy[2] + x[3] + y[3] + [4] = 0. \quad (18)$$

These equations are arranged into two 4×4 matrices containing polynomials in z:

B	xy	x	y	1
(i)	[1]	[2]	[2]	[3]
(j)	[1]	[3]	[3]	[4]
(k)	[1]	[3]	[3]	[4]
(l)	[2]	[3]	[3]	[4]

C	xy	x	y	1
(i)	[1]	[2]	[2]	[3]
(j)	[1]	[3]	[3]	[4]
(k)	[1]	[3]	[3]	[4]
(m)	[2]	[3]	[3]	[4]

Since the vector $[\ xy \quad x \quad y \quad 1\ ]^\top$ is a nullvector to both these matrices, their determinant polynomials must both vanish. Let the two eleventh degree determinant polynomials be denoted by (n) and (o) , respectively. The eleventh degree term is cancelled between them to yield the tenth degree polynomial

$$(p) \equiv (n)o_{11} - (o)n_{11}. \quad (19)$$

The real roots of (p) are now computed. There are various standard methods to accomplish this. A highly efficient way is to use Sturm-sequences to bracket the roots, followed by a root-polishing scheme. This is described in Section 3.2. Another method, which is easy to implement with most linear algebra packages, is to eigen-decompose a companion matrix. After normalising (p) so that $p_{10} = 1$, the roots are found as the eigenvalues of the 10×10 companion matrix

$$\begin{bmatrix} p_9 & p_8 & \cdots & p_0 \\ -1 & & & \\ & \ddots & & \\ & & -1 & \end{bmatrix}. \quad (20)$$

For each root z the variables x and y can be found using equation system B. The last three coordinates of a nullvector to B are computed, for example by evaluating the three 3×3 determinants obtained from the first three rows of B by striking out the columns corresponding to x, y and 1, respectively. The essential matrix is then obtained from (9). In Section 3.1 it is described how to recover R and t from the essential matrix.

3.1 Recovering R and t from E

Let

$$D = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (21)$$

R and t are recovered from the essential matrix on the basis of the following theorem [30, 10]:

Theorem 2 *Let the singular value decomposition of the essential matrix be $E \sim U diag(1, 1, 0) V^\top$, where U and V are chosen such that $det(U) > 0$ and $det(V) > 0$. Then $t \sim t_u \equiv [\ u_{13} \quad u_{23} \quad u_{33}\ ]^\top$ and R is equal to $R_a \equiv UDV^\top$ or $R_b \equiv UD^\top V^\top$.*

Any combination of R and t according to the above prescription satisfies the epipolar constraint (4). To resolve the inherent ambiguities, it is assumed that the first camera matrix is $[I \mid 0]$ and that t is of unit length. There are then the following four possible solutions for the second camera matrix: $P_A \equiv [R_a \mid t_u]$, $P_B \equiv [R_a \mid -t_u]$, $P_C \equiv [R_b \mid t_u]$, $P_D \equiv [R_b \mid -t_u]$. One of the four choices corresponds to the true configuration. Another one corresponds to the twisted pair which is obtained by rotating one of the views 180 degrees around the baseline. The remaining two correspond to reflections of the true configuration and the twisted pair. For example, P_A gives one configuration. P_C corresponds to its twisted pair, which is obtained by applying the transformation

$$H_t \equiv \begin{bmatrix} & I & & 0 \\ -2v_{13} & -2v_{23} & -2v_{33} & -1 \end{bmatrix}. \quad (22)$$

P_B and P_D correspond to the reflections obtained by applying $H_r \equiv diag(1, 1, 1, -1)$. In order to determine which choice corresponds to the true configuration, the cheirality constraint [1] is imposed. One point is sufficient to resolve the ambiguity. The point is triangulated using the view pair $([I \mid 0], P_A)$ to yield the space point Q and cheirality is tested. If $c_1 \equiv Q_3 Q_4 < 0$, the point is behind the first camera. If $c_2 \equiv (P_A Q)_3 Q_4 < 0$, the point is behind the second camera. If $c_1 > 0$ and $c_2 > 0$, P_A and Q correspond to the true configuration. If $c_1 < 0$ and $c_2 < 0$, the reflection H_r is applied and we get P_B. If on the other hand $c_1 c_2 < 0$, the twist H_t is applied and we get P_C and the point $H_t Q$. In this case, if $Q_3 (H_t Q)_4 > 0$ we are done. Otherwise, the reflection H_r is applied and we get P_D.

[1] The constraint that the scene points should be in front of the cameras.

3.2 Efficiency Considerations

In summary, the main computational steps of the algorithm outlined above are as follows:

1. Extraction of the nullspace of a 5×9 matrix.
2. Expansion of the cubic constraints (5).
3. Gauss-Jordan elimination on the 9×20 matrix A.
4. Expansion of the determinant polynomials of the two 4×4 polynomial matrices B and C followed by elimination to obtain the tenth degree polynomial (19).
5. Extraction of roots from the tenth degree polynomial.
6. Recovery of R and t corresponding to each real root and point triangulation for disambiguation.

We will discuss efficient implementation of Steps 1,5 and 6. Singular value decomposition is the gold standard for the nullspace extraction in Step 1, but a specifically tailored QR-factorisation is much more efficient. The five input vectors are orthogonalised first, while pivoting, to form the orthogonal basis $\tilde{q}_1, \ldots, \tilde{q}_5$. This basis is then amended with the 9×9 identity matrix to form the matrix

$$\left[\begin{array}{ccc|c} \tilde{q}_1 & \cdots & \tilde{q}_5 & I \end{array} \right]^\top \tag{23}$$

The orthogonalisation with pivoting is now continued until nine orthogonal vectors are obtained. The last four rows constitute an orthogonal basis for the nullspace.

Sturm sequences are used to bracket the roots in Step 5. The definition of a Sturm sequence, also called Sturm chain is given in Appendix A. The tenth degree polynomial has an associated Sturm sequence, which consists of eleven polynomials of degree zero to ten. The number of real roots in an interval can be determined by counting the number of sign changes in the Sturm sequence at the two endpoints of the interval. The Sturm sequence can be evaluated recursively with 38 floating point operations. 10 additional operations are required to count the number of sign changes. This is to be put in relation to the 20 floating point operations required to evaluate the polynomial itself. With this simple test for number of roots in an interval, it is fairly straightforward to hunt down a number of intervals, each containing one of the real roots of the polynomial. Any root polishing scheme [24] can then be used to determine the roots accurately. In our experiments we simply use 30 iterations of bisection, since this provides a guaranteed precision in fixed time and requires almost no control overhead.

Step 6 requires a singular value decomposition of the essential matrix and triangulation of one or more points. When all the other steps of the algorithm have been efficiently implemented, these operations can take a significant portion of the computation time, since they have to be carried out for each real root. A specifically tailored singular value decomposition is given in Appendix B. Efficient triangulation is discussed in Appendix C. Note that a triangulation scheme that assumes ideal point correspondences can be used since for true solutions the recovered essential matrix is such that intersection is guaranteed for the five pairs of rays.

4. Planar Structure Degeneracy

The planar structure degeneracy is an interesting example of the differences between the calibrated and uncalibrated frameworks. The degrees of ambiguity that arise from a planar scene in the two frameworks are summarised in Table 1. For pose estimation with known intrinsics there is a unique solution provided that the plane is finite and that the cheirality constraint is taken into account [2]. In theory, focal length can also be determined if the principal direction does not coincide with the plane normal. Without knowledge of the intrinsics however, there is a three degree of freedom ambiguity that can be thought of as parameterised by the position of the camera centre. For any camera centre, appropriate choices for the calibration matrix K and rotation matrix R can together produce any homography between the plane and the image. With known intrinsics and two views of an unknown plane, there are two solutions for the essential matrix [15, 17], unless the baseline is perpendicular to the plane in which case there is a unique solution. The cheirality constraint resolves the ambiguity unless all visible points are closer to one viewpoint than the other [15]. If all visible points are closer to one viewpoint, the dual solution is obtained from the true one by reflecting that view across the plane and then taking the twisted pair of the resulting configuration. Any attempts to recover intrinsic parameters from two views of a planar surface are futile according to the following theorem, adapted from [16]:

Theorem 3 *For any choice of intrinsic parameters, any homography can be realised between two views by some positioning of the two views and a plane.*

If the calibration matrices are completely unknown, there is a two degree of freedom ambiguity, that can be thought of as parameterised by the epipole in one of the images, i.e. for any choice of epipole in the first image, there is a unique valid solution. Once the epipole is specified in the first image, the problem of solving for the remaining parameters of the fundamental matrix is algebraically equivalent to solving for the projective pose of a one-dimensional camera in a two-dimensional world, where the projection centre of the

[2] If the plane is the plane at infinity it is impossible to determine the camera position and without the cheirality constraint the reflection across the plane constitutes a second solution.

	1 View Known Structure	2 Views Unknown Structure	$n > 2$ Views Unknown Structure
Known intrinsics	Unique	Two-fold or unique	Unique
Unknown fixed focal length	Unique in general	1 d.o.f.	Unique in general
Unknown variable intrinsics	3 d.o.f.	2 d.o.f.	3n-4 d.o.f.

Table 1: The degrees of ambiguity in the face of planar degeneracy for pose estimation and structure and motion estimation. The motion is assumed to be general and the structure is assumed to be dense in the plane. See the text for further explanation.

1-D camera corresponds to the epipole in the second image, the orientation corresponds to the epipolar line homography and the points in the second image correspond to world points in the 2-D space. The problem according to Chasles' Theorem [10] has a unique solution unless all the points and the epipole in the second image lie on a conic, which is not the case since we are assuming that the structure is dense in the plane. For three views with known intrinsics there is a unique solution. If the views are in general position a common unknown focal length can also be recovered, but this requires rotation and suffers from additional critical configurations. With unknown variable intrinsics there are 3 additional degrees of freedom for each view above two.

5. Applying the Algorithm Together with Random Sample Consensus

We use the algorithm in conjunction with random sampling consensus in two or three views. A number of random samples are taken, each containing five point-tracks. The five-point algorithm is applied to each sample and thus a number of hypotheses are generated. In the two-view case, the hypotheses are scored by a robust measure over all the point pairs and the hypothesis with the best score is retained. Finally, the best hypothesis can be polished by iterative refinement [29]. When three or more views are available, we prefer to disambiguate and score the hypotheses utilising three views. A unique solution can then be obtained from each sample of five tracks and this continues to hold true even if the scene points are all perfectly coplanar. For each sample of five point-tracks, the points in the first and last view are used in the five-point algorithm to determine a number of possible camera matrices for the first and last view. For each case, the five points are triangulated [3]. The remaining view can now be determined by any 3-point calibrated perspective pose algorithm, see [8] for a review and additional references. Up to four solutions are obtained and disambiguated by the additional two points. The reprojection errors of the five points in all of the views are now enough to single out one hypothesis per sample. Finally, the solutions from all samples are scored by a robust measure using all available point tracks.

[3] See Appendix C.

6. Results

For a minimal solution such as the five-point method the two main requirements are precision and speed. Observe that the effects of noise will be the same for any five-point solution method. The numerical precision of our fast implementation is investigated in Figure 1. Note that the typical errors are insignificant in comparison to realistic noise levels.

The computation time is partially dependent on the number of real solutions. The distribution of the number of solutions is given in Table 2. We have also verified experimentally that five points in three views in general yield a unique solution, with or without planar structure and an unknown focal length common to the three views. Timing information for our efficient implementation of the five-point algorithm is given in Table 3.

The algorithm is used as a part of a system that reconstructs structure and motion from video in real-time. System timing information is given in Table 4. Some results from the reconstruction system are given in Figures 2-6. See the figure captions for details.

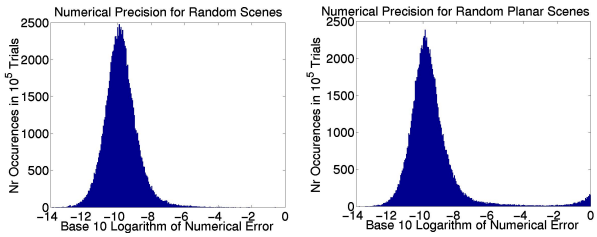

Figure 1: Distribution of the numerical error in the computed essential matrix $\hat{E}$ based on 10^5 random tests with generic (left) and planar (right) scenes. Since the matrix is up to scale and there are multiple solutions $\hat{E}_i$, the minimum residual $\min_i \min(\| \frac{\hat{E}_i}{\|\hat{E}_i\|} - \frac{E}{\|E\|} \|, \| \frac{E}{\|E\|} + \frac{\hat{E}_i}{\|\hat{E}_i\|} \|)$ from each problem instance is used. The median error is $1.39 \cdot 10^{-10}$ for generic scenes and $1.76 \cdot 10^{-10}$ for planar scenes. All computations were performed in double precision.

7. Summary and Conclusions

An efficient algorithm for solving the five-point relative pose problem was presented. The algorithm was used in conjunction with random sampling consensus to solve for unknown structure and motion over two, three or more views. The efficiency of the algorithm is very important since it will typically be applied within this kind of hypothesise-and-test architecture, where the algorithm is executed for hundreds of different five-point samples. Practical real-time reconstruction results were given and it was shown that the calibrated framework can continue to operate correctly despite scene planarity.

Nr Hyp	0	1	2	3	4	5	6	7	8	9	10
Step 5	0	.	0.12	.	0.50	.	0.36	.	0.15	.	$4.9e-4$
Step 6	$4.2e-6$	0.17	0.28	0.29	0.17	$5.8e-2$	$2.5e-2$	$1.5e-3$	$6.6e-4$	$1.5e-6$	$2e-7$

Table 2: The distribution of the number of hypotheses that result from computational steps 5 and 6 (as numbered in Section 3.2). The second row shows the distribution of the number of real roots of the tenth degree polynomial (p) in Equation (19), based on 10^5 random point and view configurations. The average is 4.55 roots. The third row shows the distribution of the number of hypotheses once the cheirality constraint has been enforced, based on 10^7 random point and view configurations. The average number of hypotheses is 2.74. Both rows show fractions of the total number of trials. Our current randomization leads to two cases in 10^7 with ten distinct physically valid solutions. We have verified that there are such cases with ten well separated solutions that are not caused by numerical inaccuracies.

Step	1	2	3	4	5	6	Three-Point Pose	Mean Two Views	Mean Three Views
μs	8	12	23	14	$6/root$	$8/root$	$5/root$	121	134

Table 3: Approximate timings for the algorithm steps (as numbered in Section 3.2) on a modest 550MHz machine with highly optimised but platform-independent code. Including all overhead, the two and three view functions typically take 110-140μs and 120-180μs, respectively. For RANSAC processes with 500 samples the total hypothesis generation times are around $60ms$ and $67ms$, respectively.

Feature Detection	Matching with Disparity Range			SaM
	3%	5%	10%	
$30ms$	$34ms$	$45ms$	$160ms$	$50ms$

Table 4: Approximate average timings per 720x240 frame of video for the system components on a modest 550MHz machine. MMX code was used for the crucial parts of the feature detection and feature matching. Disparity range for the matching is given in percent of the image dimensions. In the structure and motion component (SaM), one-view and three-view estimations are combined to incrementally build the reconstruction with low latency. The whole system including all overhead currently operates at 26 frames per second on average on a 2.4GHz machine when using a 3% disparity range. The latency is also small, since there is no self-calibration and only very local iterative refinements.

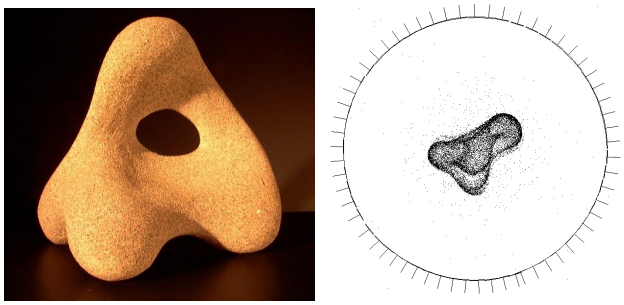

Figure 2: Reconstruction from the turntable sequence 'Stone'. No prior knowledge about the motion or the fact that it closes on itself was used in the estimation. The circular shape of the estimated trajectory is a verification of the correctness of the result. This result was obtained *without* any global bundle adjustment and exhibits a regularity and accuracy that is typically not obtained with an uncalibrated method until the calibration constraints have been enforced through bundle adjustment.

Appendixes

A Definition of Sturm Chain

Let $p(z)$ be a general polynomial of degree $n >= 2$. Here, the significance of general is that we ignore special cases for the sake of brevity. For example, $p(z)$ is assumed to have no multiple roots. Moreover, the polynomial divisions carried out below are assumed to have a non-zero remainder. Under these assumptions, the Sturm chain is a sequence of polynomials $f_0, \ldots, f_n$ of degrees $0, \ldots, n$, respectively. f_n is the polynomial itself and f_{n-1} is its derivative:

$$f_n(z) \equiv p(z) \quad (24)$$
$$f_{n-1}(z) \equiv p'(z). \quad (25)$$

For $i = n, \ldots, 2$ we carry out the polynomial division f_i/f_{i-1}. Let the quotient of this division be $q_i(z) = k_i z + m_i$ and let the remainder be $r_i(z)$, i.e. $f_i(z) = q_i(z)f_{i-1}(z) + r_i(z)$. Then define $f_{i-2}(z) \equiv -r_i(z)$. Finally, define the coefficients m_0, m_1 and k_1 such that

$$f_0(z) = m_0 \quad (26)$$
$$f_1(z) = k_1 z + m_1. \quad (27)$$

Once the scalar coefficients $k_1, \ldots, k_n$ and $m_0, \ldots, m_n$ have been derived, the Sturm chain can be evaluated at any point z through Equations (26, 27) and the recursion

$$f_i(z) = (k_i z + m_i) f_{i-1}(z) - f_{i-2}(z) \quad i = 2, \ldots, n \quad (28)$$

Let the number of sign changes in the chain be $s(z)$. The number of real roots in an interval $[a, b]$ is then $s(a) - s(b)$. Unbounded intervals such as for example $[0, \infty)$ can be treated by looking at m_0 and $k_0, \ldots, k_n$ in order to calculate $\lim_{z\to\infty} s(z)$. For more details, see for example [7, 12].

B Efficient Singular Value Decomposition of the Essential Matrix

An efficient singular value decomposition according to the conditions of Theorem 2 is given. Let the essential matrix be $E = \begin{bmatrix} e_a & e_b & e_c \end{bmatrix}^\top$, where e_a, e_b, e_c are column-vectors. It is assumed that it is a true essential matrix, i.e. that it has rank two and two equal non-zero singular values. First, all the vector products $e_a \times e_b$, $e_a \times e_c$ and $e_b \times e_c$ are computed and the one with the largest magnitude chosen.

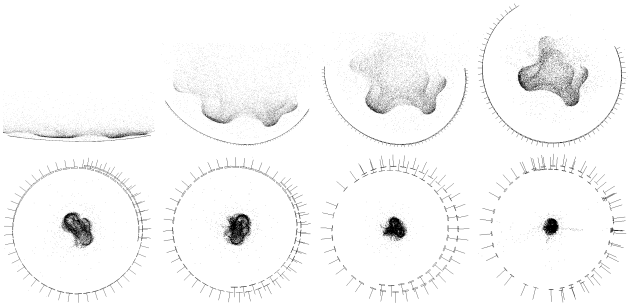

Figure 3: Reconstructions obtained from the 'Stone' sequence by setting the focal length to incorrect values. The focal lengths used were 0.05, 0.3, 0.5, 0.7, 1.3, 1.5, 2.0 and 3.0 times the value obtained from calibration. For too small focal lengths, the reconstruction 'unfolds' and vice versa.

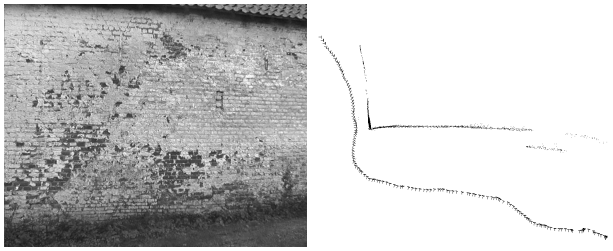

Figure 4: Reconstruction from the sequence 'Farmhouse', which contains long portions where a single plane fills the field of view. The successful reconstruction is a strong practical proof of the fact that the calibrated framework can overcome planar structure degeneracy without relying on the degeneracy or trying to detect it. This is especially important for near-planar scenes, where neither the planar nor the uncalibrated model applies well. Only approximate intrinsic parameters were used and no global bundle adjustment was performed.

Assume without loss of generality that $e_a \times e_b$ has the largest magnitude. Define $v_c \equiv (e_a \times e_b)/|e_a \times e_b|$, $v_a \equiv e_a/|e_a|$, $v_b \equiv v_c \times v_a$, $u_a \equiv Ev_a/|Ev_a|$, $u_b \equiv Ev_b/|Ev_b|$ and $u_c \equiv u_a \times u_b$. Then the singular value decomposition is given by $V = \left[\begin{array}{ccc} v_a & v_b & v_c \end{array} \right]$ and $U = \left[\begin{array}{ccc} u_a & u_b & u_c \end{array} \right]$.

C Efficient Triangulation of an Ideal Point Correspondence

In the situation encountered in the five-point algorithm where triangulation is needed, a hypothesis for the essential matrix E has been recovered and along with it the two camera matrices $[I \mid 0]$ and P. No error metric has to be minimised, since for the true solution the rays backprojected from the image correspondence $q \leftrightarrow q'$ are guaranteed to meet. For non-ideal point correspondences, prior correction to guarantee ray-intersection while minimising a good error metric is recommended. Global minimisation of $\|.\|_2$-norm in two views requires solving a sixth degree

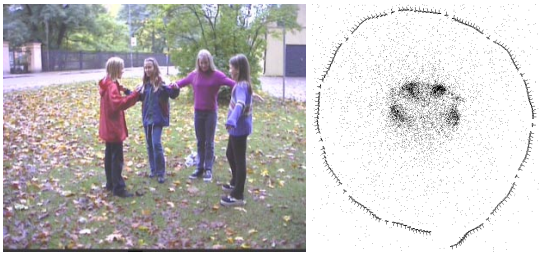

Figure 5: Reconstruction from the sequence 'Girlsstatue' that was acquired with a handheld camera. Only approximate intrinsic parameters were used and no global bundle adjustment was performed.

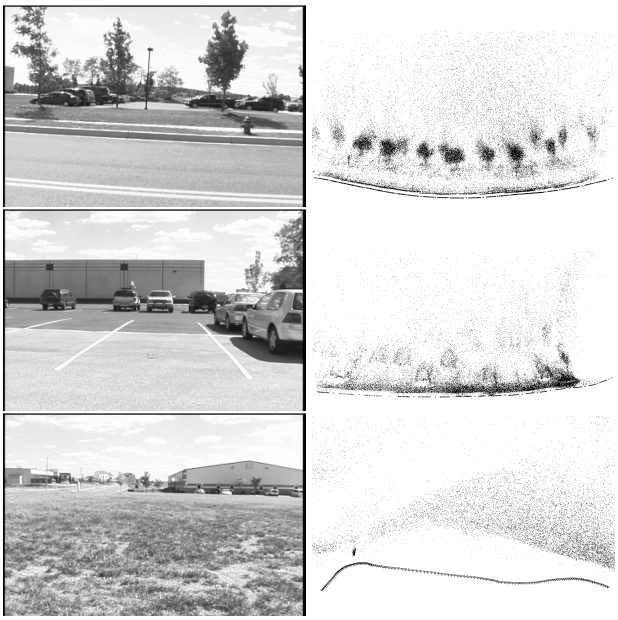

Figure 6: Reconstruction from vehicle sequences 'Road', 'Parking Lot' and 'Turn', with 360, 330 and 130 frames, respectively. Only approximate intrinsic parameters were used and no global bundle adjustment was performed.

polynomial, see [10]. Minimisation of $\|.\|_\infty$-norm [19], or directional error [20], also yields good results in practice and can be achieved in closed form an order of magnitude faster. In the ideal situation, triangulation can be accomplished very efficiently by intersecting three planes that are back-projected from image lines. The image lines chosen to generate the three planes are the epipolar line a corresponding to q', the line b through q that is perpendicular to a and the line c through q' that is perpendicular to Eq. For non-ideal point correspondences, this scheme finds the world point on the ray backprojected from q' that minimises the reprojection error in the first image. It triangulates world points at infinity correctly and is invariant to projective transformations of the world space. Observe that $a = E^\top q'$, $b = q \times (diag(1,1,0)a)$ and $c = q' \times (diag(1,1,0)Eq)$. Moreover, $A \equiv \left[\begin{array}{cc} a^\top & 0 \end{array} \right]^\top$

is the plane backprojected from a, $B \equiv [\ b^\top \quad 0\ ]^\top$ is the plane backprojected from b and $C \equiv P^\top c$ is the plane backprojected from c. The intersection between the three planes A, B and C is now sought. Formally, the intersection is the contraction $Q_l \equiv \epsilon_{ijkl} A^i B^j C^k$ between the epsilon tensor ϵ_{ijkl} [4] and the three planes. More concretely, $d \equiv a \times b$ is the direction of the ray backprojected from the intersection between a and b. The space point is the intersection between this ray and the plane C:

$$Q \sim [\ d^\top C_4 \quad -(d_1 C_1 + d_2 C_2 + d_3 C_3)\ ]^\top . \qquad (29)$$

Finally, it is observed that in the particular case of an ideal point correspondence we have $d = q$, so that computing a, b and A, B can be avoided altogether.

References

[1] P. Beardsley, A. Zisserman and D. Murray, Sequential updating of projective and affine structure from motion, *International Journal of Computer Vision*, 23(3): 235-259, 1997.

[2] M. Demazure, Sur Deux Problemes de Reconstruction, *Technical Report No 882, INRIA, Rocquencourt, France*, 1988.

[3] O. Faugeras and S. Maybank, Motion from Point Matches: Multiplicity of Solutions, *International Journal of Computer Vision*, 4(3):225-246, 1990.

[4] O. Faugeras, What Can be Seen in Three Dimensions with an Uncalibrated Stereo Rig?, *Proc. European Conference on Computer Vision*, pp. 563-578, 1992.

[5] O. Faugeras, *Three-Dimensional Computer Vision: a Geometric Viewpoint*, MIT Press, ISBN 0-262-06158-9, 1993.

[6] M. Fischler and R. Bolles, Random Sample Consensus: a Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography, *Commun. Assoc. Comp. Mach.*, 24:381-395, 1981.

[7] W. Gellert, K. Küstner, M. Hellwich and H. Kästner, *The VNR Concise Encyclopedia of Mathematics*, Van Nostrand Reinhold Company, ISBN 0-442-22646-2, 1975.

[8] R. Haralick, C. Lee, K. Ottenberg and M. Nölle, Review and Analysis of Solutions of the Three Point Perspective Pose Estimation Problem, *International Journal of Computer Vision*, 13(3):331-356, 1994.

[9] R. Hartley, Estimation of Relative Camera Positions for Uncalibrated Cameras, *Proc. European Conference on Computer Vision*, pp. 579-587, 1992.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN 0-521-62304-9, 2000.

[11] A. Heyden and G. Sparr, Reconstruction from Calibrated Cameras - a New Proof of the Kruppa-Demazure Theorem, *Journal of Mathematical Imaging & Vision*, 10:1-20, 1999.

[12] D. Hook and P. McAree, Using Sturm Sequences To Bracket Real Roots of Polynomial Equations, *Graphic Gems I*, Academic Press, ISBN 0-122-86166-3, pp. 416-423, 1990.

[13] B. Horn, Relative Orientation, *International Journal of Computer Vision*, 4:59-78, 1990.

[14] E. Kruppa, Zur Ermittlung eines Objektes aus zwei Perspektiven mit Innerer Orientierung, *Sitz.-Ber. Akad. Wiss., Wien, Math. Naturw. Kl., Abt. IIa.*, 122:1939-1948, 1913.

[15] H. Longuet-Higgins, The Reconstruction of a Plane Surface from Two Perspective Projections, *Proc. R. Soc. Lond. B*, 277:399-410, 1986.

[16] S. Maybank, *Theory of Reconstruction from Image Motion*, Springer-Verlag, ISBN 3-540-55537-4, 1993.

[17] S. Negahdaripour, Closed-Form Relationship Between the Two Interpretations of a Moving Plane, *J. Optical Society of America*, 7(2):279-285, 1990.

[18] D. Nistér. Reconstruction From Uncalibrated Sequences with a Hierarchy of Trifocal Tensors, *Proc. European Conference on Computer Vision*, Volume 1, pp. 649-663, 2000.

[19] D. Nistér. *Automatic dense reconstruction from uncalibrated video sequences*, PhD Thesis, Royal Institute of Technology KTH, ISBN 91-7283-053-0, March 2001.

[20] J. Oliensis and Y. Genc, New Algorithms for Two-Frame Structure from Motion, *Proc. International Conference on Computer Vision*, pp. 737-744 ,1999.

[21] J. Philip, A Non-Iterative Algorithm for Determining all Essential Matrices Corresponding to Five Point Pairs, *Photogrammetric Record*, 15(88):589-599, October 1996.

[22] M. Pollefeys, R. Koch and L. Van Gool, Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters, *International Journal of Computer Vision*, 32(1):7-25, 1999.

[23] M. Pollefeys, F. Verbiest and L. Van Gool, Surviving Dominant Planes in Uncalibrated Structure and Motion Recovery, *Proc. European Conference on Computer Vision*, Volume 2, pp. 837-851, 2002.

[24] W. Press, S. Teukolsky, W. Vetterling and B. Flannery, *Numerical recipes in C*, Cambridge University Press, ISBN 0-521-43108-5, 1988.

[25] P. Stefanovic, Relative Orientation - a New Approach, *I. T. C. Journal*, 1973-3:417-448, 1973.

[26] P. Torr and D. Murray, The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix, *International Journal of Computer Vision*, 24(3):271-300, 1997.

[27] P. Torr, A. Fitzgibbon and A. Zisserman, The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences, *International Journal of Computer Vision*, 32(1):27-44, August 1999.

[28] B. Triggs, Routines for Relative Pose of Two Calibrated Cameras from 5 Points, *Technical Report, http://www.inrialpes.fr/movi/people/Triggs INRIA, France*, 2000.

[29] B. Triggs, P. McLauchlan, R. Hartley and A. Fitzgibbon, Bundle Adjustment - a Modern Synthesis, *Springer Lecture Notes on Computer Science*, Springer Verlag, 1883:298-375, 2000.

[30] R. Tsai and T. Huang, Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):13-27, 1984.

[31] Z. Zhang, Determining the Epipolar Geometry and its Uncertainty: a Review, *International Journal of Computer Vision*, 27(2):161-195, 1998.

[4] The epsilon tensor ϵ_{ijkl} is the tensor such that $\epsilon_{ijkl} A^i B^j C^k D^l = det([\ A \quad B \quad C \quad D\ ])$.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U. S. Government.