Automatic Camera Recovery for Closed or Open Image Sequences

Andrew W. Fitzgibbon and Andrew Zisserman

Robotics Research Group, Department of Engineering Science,
University of Oxford, 19 Parks Road, Oxford OX1 3PJ, United Kingdom
{awf,az}@robots.ox.ac.uk

Abstract. We describe progress in completely automatically recovering 3D scene structure together with 3D camera positions from a sequence of images acquired by an unknown camera undergoing unknown movement.

The main departure from previous structure from motion strategies is that processing is *not* sequential. Instead a hierarchical approach is employed building from image triplets and associated trifocal tensors. This is advantageous both in obtaining correspondences and also in optimally distributing error over the sequence.

The major step forward is that closed sequences can now be dealt with easily. That is, sequences where part of a scene is revisited at a later stage in the sequence. Such sequences contain additional constraints, compared to open sequences, from which the reconstruction can now benefit.

The computed cameras and structure are the backbone of a system to build texture mapped graphical models directly from image sequences.

1 Introduction

The goal of this work is to obtain camera projection matrices and 3D structure from long sequences of uncalibrated images. Once obtained the cameras and structure are the basis for building 3D graphical models directly from images. This competence is also required for many other structure-from-motion applications, for example ego-motion determination.

There are two main aspects. The first is establishing corresponding image tokens (corners here) over all the images. This problem is exasperated because a corner feature will generally not appear in all of the images, and often will be missing from consecutive images. Sequential matchers have proved the most successful [1, 2, 3, 4, 13, 26, 27, 33].

The second aspect is distributing camera and structure "error" in an optimal manner over all the images. Optimal here is defined as minimizing the reprojection error over the sequence. In the case of affine cameras (e.g. weak perspective) the factorization method of Tomasi and Kanade [26] is optimal [18]. In the case of general perspective (which is the only case considered from here on) factorization-like methods have been developed [9, 22, 25] but these minimize an algebraic error rather than reprojection error. Furthermore, all factorization methods are limited to features for which there are correspondences in every

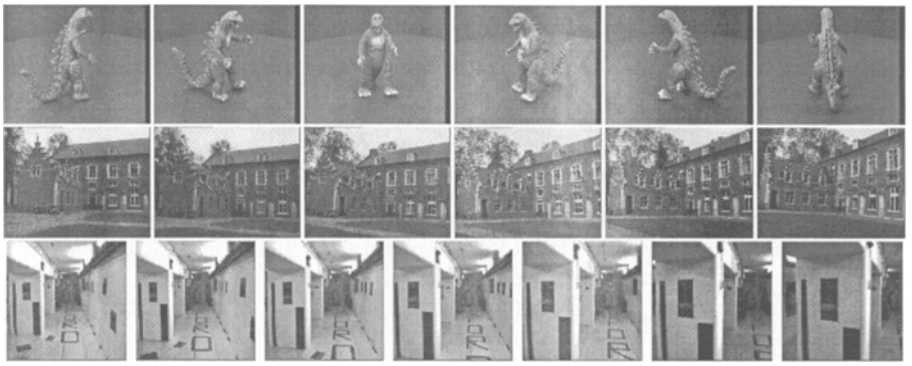


Fig. 1. Example sequences: Dinosaur on turntable (36 frames); Castle, hand-held camera (25 frames); Basement, camera on AGV (12 frames).

image, a problem that is addressed in [10]. Bundle adjustment [6, 15] involves varying the structure and cameras in order to minimize reprojection error. This is optimal, and is not hindered by missing correspondences. However, bundle adjustment does not have a direct solution (such as the SVD solution in factorization) and involves a non-linear optimization which requires a good starting point. A sub-optimal alternative to bundle adjustment is a recursive (Kalman like) filter in sequential processing [1, 3, 4, 14]. However a poor two view or three view structure initialization severely affects the accuracy of subsequent camera and structure recovery for sequential systems. The problem of distributing error over many measurements of a 3D scene is a recurring one in computer vision. Ikeuchi [21] dealt with it for range images, and Porrill [17] for a calibrated stereo head.

The starting point for the work described here is the competence in computing multiple view relations for consecutive frames of a sequence — the fundamental matrix F for image pairs and trifocal tensor [7, 20, 23] $\mathcal{T}$ for image triplets can be computed *well*. Their computation is automatic and reliable, and the estimated tensor is extremely accurate. The key idea here is to always build on this competence. As will be demonstrated in the sequel, building on this strength allows the detection and avoidance of many of the problems that plague both sequential matching and the initialization of bundle adjustment.

The building block used is an image triplet. A triplet consists of three elements: image corner correspondences between the three views; the trifocal tensor for the triplet; and the 3D structure in an arbitrary projective frame defined by three camera matrices consistent with the trifocal tensor. This basic unit is optimal in the sense that the projection matrices and 3D structure are refined by bundle adjustment. Its computation is described in section 2. Using triplets as the building block confers a number of important advantages: the strong geometric constraint means that very few false corner correspondences are encountered, line matches can also be included, and the problem of critical surfaces is reduced [12].

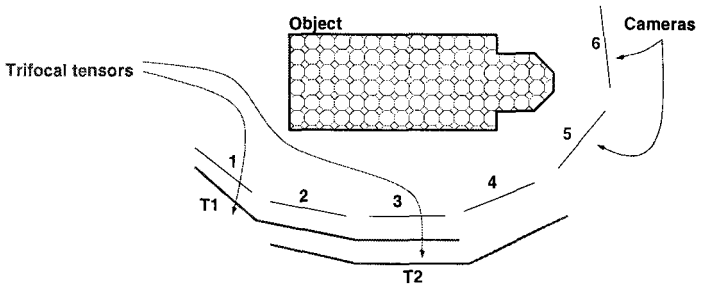


Fig. 2. Overview of the problem. A rigid object is observed by unknown cameras. Camera positions and the structure of the object are computed automatically using image triplets as the basic building block. The triplets/trifocal tensors are registered into sub-sequences.

To proceed from triplets to a complete description of a sequence it is necessary to *register* all the triplets into the same coordinate frame. One possible strategy is to register triplets directly and thence build up common 3D structure and consistent camera matrices for all views. It will be shown that there are significant advantages in instead using a hierarchical approach where triplets are registered into sub-sequences, (figure 2), followed by registering sub-sequences into the entire sequence. The overall process is summarised in figure 3. Of course, if the sub-sequence is the entire sequence then triplets are registered directly. Similar approaches to ours have been developed by Laveau [11] and Sturm [24].

1. *Optimally estimate trifocal tensors* $\mathcal{T}$ *for all consecutive image triplets.*
2. *Compute structure points* $\{\mathbf{X}_i\}$ *and camera positions* $\{\mathtt{P}_1, \mathtt{P}_2, \mathtt{P}_3\}$ *for each triplet.*
3. *Register triplets into consistent sub-sequences using 3-space homographies* $\mathtt{H}$.
4. *Bundle adjust sub-sequences.*
5. *Optional hierarchical registration of sub-sequences into longer sub-sequences.*
6. *Register sub-sequences, again using homgraphies, to obtain cameras and structure for the complete sequence.*
7. *Bundle adjust the cameras and 3D structure for the complete sequence.*

Fig. 3. Algorithm overview: Hierarchically compute correspondences, cameras and 3D structure over a sequence.

The advantages conferred by proceeding in this hierarchical fashion, as opposed to a sequential approach, are five fold, and are the main contributions

of this paper. First, erroneous processing in particular sub-sequences can be dealt with locally before the frames are combined; second, the dependency on a good estimate from the early frames of the sequence is reduced; third, the overall process can sometimes be more computationally efficient; fourth, the reconstruction after bundle adjustment is more accurate because a closer starting point is provided; finally, closed sequences can be processed easily in this framework.

Registration of triplets and sub-sequences is achieved by computing the homography of 3-space which results in the best overlap (where "best" will be defined later) of the two projective structures. Section 3 describes and compares a number of strategies for determining this homography, and the registration of triplets. Section 4 describes the registration of sub-sequences. Finally, appendix A overviews the bundle adjustment algorithm which is used at a number of stages throughout this work.

In order to be able to visually assess the reconstruction quality all cameras and 3D structure are Euclidean corrected. The auto-calibration method for this Euclidean correction proceeds from the computed camera matrices, and is based on the dual of the absolute conic parametrization of Triggs [31] together with the algorithm of Pollefeys *et al.* [16].

2 Estimation of $\mathcal{T}$ for image triplets

The foundation of the methods described in this paper is the ability to automatically and reliably compute an accurate trifocal tensor for consecutive frames of a sequence. Trifocal tensor computation has greatly improved over the computational method described in [2]. This improvement is not due solely to one factor, but to a combination of many incremental changes. As $\mathcal{T}$ estimation is not the main contribution of this paper the algorithm will not be described in detail, but the important incremental improvements are summarised.

Briefly, putative point matches (Harris corners[5]) are first obtained for the consecutive image pairs, one/two and two/three, by simultaneously computing epipolar geometry and matches consistent with this estimated geometry using a robust estimation algorithm. This is now fairly standard [2, 28, 32]. From these seed matches the trifocal tensor is robustly fitted, and new matches are found (*guided matching*) which are consistent with the fitted $\mathcal{T}$. Fitting and guided matching are repeated until the number of matched points stabilises. The improvements over [2] include:

1. Parametrizing the trifocal tensor such that it obeys all the constraints between the tensor elements [30].
2. Maximum-Likelihood Estimation (MLE) of $\mathcal{T}$ via bundle adjustment (appendix A). The use of a cost function which corresponds to reprojection error [30] rather than transfer error is one of the most important improvements, as the transfer error tends to accept points which are large outliers to the MLE distance and vice-versa.

3. Point pairs are transferred for guided matching by first Hartley-Sturm [8] correcting the pair. This ensures that the guided matching stage does not lose inliers.
4. Point triplets are accepted during guided matching based on reprojection error rather than transfer distance. This is also an important modification: if the guided matching stage uses a different distance measure to the fitting stage, the number of matches oscillates with each iteration and heuristic termination criteria are needed to decide when to stop the procedure. When the same error is used for both stages, the number of matches increases monotonically until convergence.
5. The RANSAC procedure uses the adaptive termination criterion detailed in [29, p. 286] which allows safe early termination for simple scenes without prejudicing more complex ones.

Together, these modifications result in more accurate tensors with more inliers and fewer false matches. Although some of the modifications appear expensive, the improvement in convergence properties means that the total time to estimate $\mathcal{T}$ is substantially reduced.

3 From triplets to sub-sequences

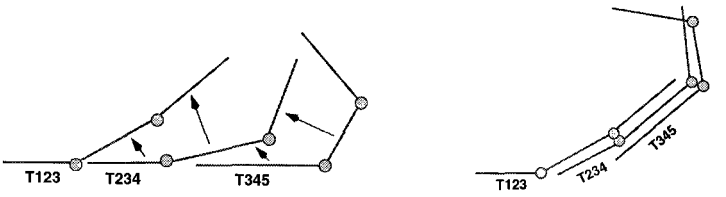


Fig. 4. Registration of trifocal tensors into consistent projective frames. Homographies of $\mathcal{P}^3$ are computed which place tensors $\mathcal{T}_{234}$ and $\mathcal{T}_{345}$ in the frame of $\mathcal{T}_{123}$.

This section describes the registration of image triplets to form sub-sequences (see figure 4). To be specific the case of registering two triplets will be considered, but the issues that arise are common to all the registration problems in the following sections.

We are given a pair of image triplets, each with an estimated trifocal tensor and a set of 3D points corresponding to image points in all views of the triplet. It is assumed that some of the 3D points are common to both sets. The goal is to obtain a common set of 3D points and a camera for each view, such that the reprojection error is minimized.

In more detail we have a set of 3D points represented in the two projective frames provided by the trifocal tensor of each triplet. Suppose a point has

coordinates $\mathbf{X}_i$ in the first triplet and $\mathbf{X}'_i$ in the second. If all measurements were perfect then there would exist a homography $\mathtt{H}$ of 3-space between the two projective frames such that

$$\mathtt{P}_j = \mathtt{P}'_j\mathtt{H}^{-1} \tag{1}$$

$$\mathbf{X}_i = \mathtt{H}\mathbf{X}'_i \tag{2}$$

where $\mathtt{P}_j, \mathtt{P}'_j$ are the corresponding camera matrices for images common to the triplets. Of course, with real image sequences the relationship will not be obeyed exactly, and an error-minimizing estimate must be found.

Registration will always proceed in two steps: first, a homography of 3 space is computed which approximately registers the triplets; second, an optimal registration is obtained by bundle adjustment (see appendix A). The different strategies are targetted on how best to obtain the approximate homography.

Zisserman *et al.* [34] and Laveau [11, §3.4] describe two ways in which this homography may be obtained. The first is to minimize a "distance" between the 3D points:

$$\min_{\mathtt{H}} \epsilon_{\mathrm{D}} = \sum_i D^2(\mathbf{X}_i, \mathtt{H}\mathbf{X}'_i) \tag{3}$$

where the distance $D(\cdot,\cdot)$ is either *algebraic* distance

$$D_A(\mathbf{X},\mathbf{Y}) = \sum_{k=1}^{3}(X_kY_4 - Y_kX_4)^2 \tag{4}$$

or *Euclidean* distance

$$D_E(\mathbf{X},\mathbf{Y}) = \sum_{k=1}^{3}\left(\frac{X_k}{X_4} - \frac{Y_k}{Y_4}\right)^2 \tag{5}$$

which is only strictly meaningful if the 3D projective frame has been corrected to metric. The second estimator minimizes *reprojection error* to the original corners from which the 3D points were triangulated

$$\min_{\mathtt{H}} \epsilon_{\mathrm{d}} = \sum_{ij} d^2(\mathtt{P}_j\mathtt{H}\mathbf{X}'_i, \mathbf{x}_{ij}) + d^2(\mathtt{P}'_j\mathtt{H}^{-1}\mathbf{X}_i, \mathbf{x}_{ij}) \tag{6}$$

where $d(\mathbf{x},\mathbf{y})$ is a Euclidean image distance between the inhomogeneous points corresponding to $\mathbf{x}$ and $\mathbf{y}$.

3.1 The degree of overlap and establishing correspondences

Two triplets can share zero, one or two images so that after registration the sub-sequences consist of six, five or four images respectively. We start with the correspondence problem when there is no overlap. Suppose, I, the last image of triplet one is the neighbour of I', the first image of triplet two. Then correspondences can be computed by simultaneously estimating $\mathtt{F}$ and corner matches

consistent with F between images I and I'. The corners in I will index some points in the 3D structure $\{\mathbf{X}_i\}$ of triplet one, and corners in I' will index some points in $\{\mathbf{X}'_i\}$. Thus the corner correspondence provides a partial correspondence between the two 3D point sets $\{\mathbf{X}_i\}$ and $\{\mathbf{X}'_i\}$.

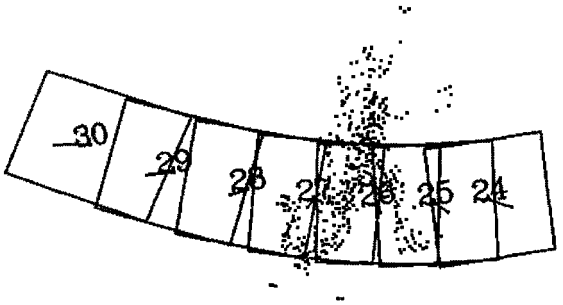


Fig. 5. Registered subsequence. Registered cameras and point structure for a 7-frame sub-sequence of the Dinosaur sequence. The cameras (numbered) are represented by their image planes and principal axes.

A much simpler solution has been suggested by Laveau [11]. Suppose there is an overlap of one view. Then image I is image I'. Consequently corner points in I/I' directly supply partial correspondences between the two 3D point sets $\{\mathbf{X}_i\}$ and $\{\mathbf{X}'_i\}$. The correspondence is only partial because a 3D point is only instantiated in a triplet if there are corner correspondences in all 3 images. So a particular corner might have an associated 3D point in one triplet but not the other. An analogous simplification in establishing correspondences applies if there is a two view overlap.

3.2 Obtaining the initial homography

We now describe methods of computing H which allow direct solutions (such as SVD). These solutions generally minimize an algebraic error with no direct geometric or statistical meaning. Furthermore, the solutions are not covariant with the choice of projective frame. However these methods are necessary for initialization of the bundle adjustment that follows, and of interest in themselves. In some cases the direct solution can be further refined by a nonlinear stage (to use reprojection error rather than algebraic distance for example) before the final bundle adjustment.

Method I: Direct 3D point registration The 3D algebraic distance (4) is readily minimized using linear algebraic methods. The Euclidean distance (5) can be solved in closed form when H is limited to similarity transformations, but requires a nonlinear minimization when the homography is allowed to be a general projectivity. This method is applicable for any number of overlapping views, including zero.

Method II: Enforcing camera consistency: one-view overlap Since a homography has 15 degrees of freedom and a camera matrix only 11, two cameras $\mathtt{P}$ and $\mathtt{P}'$ can be exactly registered, by (1). This constrains all but four of the parameters of $\mathtt{H}$. The determination of the remaining parameters will now be described.

We seek a homography $\mathtt{H}$ which minimizes $\epsilon_{\mathrm{D}} = \sum_i D^2(\mathbf{X}_i, \mathtt{H}\mathbf{X}'_i)$ subject to the constraint $\mathtt{PH} = \mathtt{P}'$. The solution is a member of the 4-parameter family of homographies:

$$\mathtt{H}(\mathbf{v}) = \mathtt{P}^{+}\mathtt{P}' + \mathbf{h}\mathbf{v}^{\top}$$

where $\mathbf{h}$ is the nullvector and $\mathtt{P}^{+}$ the pseudoinverse of $\mathtt{P}$. When D is the algebraic distance, a direct solution for $\mathbf{v}$ is obtained, leading to a system of 3 equations for $\mathbf{v}$ per 3D point:

$$\mathbf{b}\mathbf{X}'^{\top}\mathbf{v} = \mathbf{c}$$

where the 3-vectors $\mathbf{b}$ and $\mathbf{c}$ are defined by $b_k = h_k X_4 - h_4 X_k$, $c_k = X_k a_4 - X_4 a_k$ and $\mathbf{a} = \mathtt{P}^{+}\mathtt{P}'\mathbf{X}'$. When D is Euclidean distance or reprojection error, a nonlinear minimization over $\mathbf{v}$ is required.

Method III: Maximizing camera consistency: two-view overlap In this case, there is a pair of overlapping projection matrices, say $\mathtt{P}_1, \mathtt{P}_2$ which overlap with $\mathtt{P}'_1, \mathtt{P}'_2$. Each overlapping matrix $\mathtt{P}_i, \mathtt{P}'_i$ provides 11 linear constraints on the elements of $\mathtt{H}$ and therefore two or more overlapping views are sufficient to over determine $\mathtt{H}$ in (1), without recourse to the 3D point information. In general, this process is followed by a minimization of reprojection error before proceeding to bundle adjustment.

3.3 Comparison of triplet registration methods

In this section we compare registration methods in terms of accuracy, reliability and computational cost. First the one-view and two-view methods are compared and then experimental results are provided.

One-view overlap

Update based on single-view overlap is fast because trifocal tensors are required only for triplets which begin every second image. Conversely, however, it is not clear how one might use the alternate triplets. For example, having combined triplets 123, 345 and 567, how can use be made of triplets 234 and 456? On the other hand, the one-view version requires that 3D point matches be available, which implies that some feature tracking must be maintained for five images. Although this imposes stringent robustness requirements on the trifocal tensor computation, it has not proved a problem in the hundreds of images on which the algorithm has been tested. Typically 400 corners are extracted from each image which yields on average 50 to 100 correspondences over the five views. On the positive side, such tracks correspond to the most reliably located and detected 3D points, so that registration accuracy is maintained. Of course, the

Abbreviation	Overlap	Algorithm
lin	1 view	II, linear algebraic
linc	1 view	"lin", conditioned
euc	1 view	II, nonlinear 3D distance
eucc	1 view	"euc", conditioned
2view	2 views	III + Reprojection error

Sequence	Algorithm	Initial Error	Final Error	Iterations
	lin	0.319	0.198	9
	linc	0.323	0.198	8
	euc	0.329	0.198	20
	eucc	0.336	0.198	8
	2view	0.357	0.205	8
	lin	0.417	0.197	10
	linc	0.574	0.197	23
	euc	0.538	0.197	25
	eucc	0.516	0.197	23
	2view	0.309	0.215	18
	lin	0.261	0.197	18
	linc	0.264	0.197	18
	euc	0.262	0.197	18
	eucc	0.263	0.197	18
	2view	0.366	0.225	20
	lin	0.773	0.229	19
	linc	0.562	0.228	17
	euc	0.627	0.224	24
	eucc	0.441	0.226	17
	2view	0.482	0.244	16
	lin	1.213	0.461	10
	linc	1.218	0.461	22
	euc	1.209	0.461	10
	eucc	1.213	0.461	22
	2view	0.920	0.124	23

Table 1. Comparison of triplet registration algorithms. The first three tables are for subsequences 0-6, 6-12 and 12-18 of the Dinosaur sequence. The others are for the first seven frames of the Basement and Castle sequences. The initial and final error columns are the (average) reprojection errors in pixels before and after bundle adjustment. The iterations column shows the number of bundle adjustment steps required for each algorithm.

final bundle adjustment uses all correspondences (not just the 5-view ones) to obtain greater accuracy.

If using Method I, at least five 3D point correspondences are required to constrain H, and many more are needed to obtain a reliable least squares estimate. Using Method II, just two correspondences are required, meaning that the estimate is more reliable, 50 corresponding points providing a reasonable basis for the least squares computation.

Two-view overlap

With two-view overlap 3D points are not required, at least for the linear algorithm. Therefore there is no dependence on the distance metric in 3D. However, it is again difficult both to interpret the measure being minimized in terms of maximum likelihood estimation and indeed to define a ML estimator for H without recourse to the 3D point information. The two-view overlap has the advantage that shorter tracks are needed—four views rather than five—and that any false inliers to the trifocal tensors may be identified because their tracks are inconsistent. For example, a 3D point in the triplet $(1-2-3)$ may be matched to image corners numbered $(100, 200, 300)$, say, in each frame. If the corner match $(200, p, q)$ appears in the second triplet and $p \neq 300$, then one of the two triplet matches is incorrect. In the current implementation, the absence of further information means that both should be rejected. Finally, two view overlap means that all tensors in the sequence are used, which can lead to an improvement in accuracy at the expense of computational effort.

Experimental results

Table 1 compares the triplet registration strategies on a number of sequences. Each strategy was used to sequentially register seven images, and the RMS reprojection error before and after bundle adjustment was recorded. Seven images were registered in order to reflect the use to which they will be put in the following sections. An example of the registered views and structure is shown in figure 5.

The algorithms were compared using the direct linear algorithms, in both preconditioned versions (for which the 3D points were centered on the unit cube prior to computation) and non-preconditioned versions. Execution time for all algorithms is very similar: about one tenth of a second for 200 points on a Sun Ultra 170. Non-linear minimization is carried out using the Levenberg-Marquardt algorithm.

The table shows that the one-view algorithms all perform similarly, with the linear algorithm the cheapest (in terms of the number of bundle-adjustment steps subsequently needed) on high-quality laboratory sequences. However, the best all-round choice is the conditioned nonlinear algorithm. The performance of the two-view algorithm is more variable than the single view techniques, but in some cases it can lead to significantly better results. In the light of this, the recommended strategy is to use both approaches and select whichever yields

lower reprojection error for each sub-sequence. If speed is more important than reliability, the single-view linear approach is indicated.

4 From sub-sequences to sequences

The previous section describes and compares methods for the registration of triplets. Here we describe the subsequence pasting. Clearly the triplet registration approaches extend perfectly naturally to the registration of longer subsequences, being defined purely in terms of the camera projection matrices and 3D points. This can be extended to the entire input sequence, giving an algorithm similar in spirit to the earlier purely sequential approaches. On occasion however, a more hierarchical approach can yield superior results. One such situation occurs when the sequence is *closed*, visiting the same points of the object at the beginning and end of (or at any other times during) the sequence. Even with an open sequence, the hierarchical approach confers advantages in terms of speed and accuracy.

4.1 Closed sequences

In the case of a closed sequence, for example if the camera completely circumnavigates an object, there is a very tight constraint available — if there is a single overlapping frame then the camera at the head of the sequence coincides with the camera at the tail. The extent to which the computed cameras differ is a clear measure of the success of the camera recovery.

Furthermore, by explicitly enforcing the constraint that the cameras are the same, a significant improvement in accuracy is obtained. Although bundle adjustment should theoretically be able to distribute the error in such a way as to close the sequence, the point from which it starts is often sufficiently far from the true solution that the bundle adjustment converges to a local minimum.

Using the sub-sequence registration paradigm, we can solve this problem easily. Conceptually we break the full sequence into sub-sequences which are then "hinged" together using homographies. Each subsequence is then allowed to deform under a homography in order to minimize the 3D error. Because the entire system can be transformed by a homography without changing the relative positioning of the sub-sequences, one sub-sequence is chosen as a basis and remains unchanged through the minimization. Taking as an example the case of four sub-sequences numbered 1 through 4, with sub-sequence 1 as the basis, the error to be minimized is

$$\sum D^2(\mathbf{X}_1, \mathtt{H}_2\mathbf{X}_2) + \sum D^2(\mathtt{H}_2\mathbf{X}_2, \mathtt{H}_3\mathbf{X}_3) + \sum D^2(\mathtt{H}_3\mathbf{X}_3, \mathtt{H}_4\mathbf{X}_4) + \sum D^2(\mathtt{H}_4\mathbf{X}_4, \mathbf{X}_1)$$

where the sums are taken over the overlapping subsets of the 3D points in each frame. With this error function the choice of basis sequence will affect the result. If instead reprojection error is used, the basis is immaterial as the error function is invariant to homographies of space. However an advantage of the 3-space error

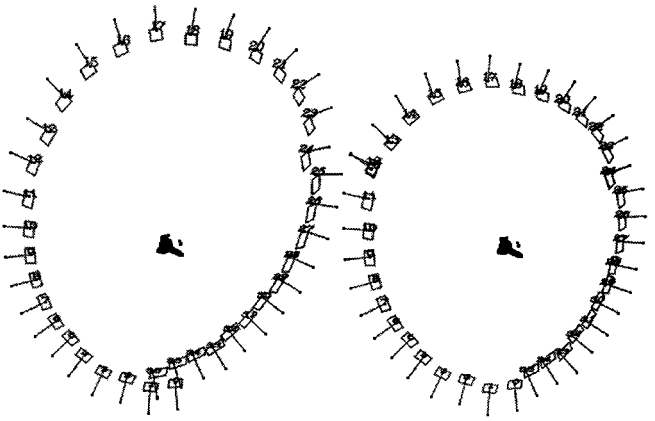


Fig. 6. Closed sequence before and after imposition of closure constraint. The left figure shows a plan view of the cameras for the Dinosaur scene as recovered by the hierarchical process. The right figure shows the same cameras after closure. Frames 0-12 are in exactly the same position in both models. The double cameras at positions 12 and 24 show where the structure was "hinged".

distance is that the camera centres may be included as part of the overlapping point sets, which assists the process when the cameras are far from the object as they are in the dinosaur sequence.

Solving the closure problem in this way has the effect of distributing the closure constraint over the entire sequence in an approximate—but computationally tractable—manner. A final bundle adjustment completes the procedure. Figure 6 illustrates the efficacy of this approach on a sequence taken using a rotating turntable.

4.2 Open sequences

Although the above technique was described as a solution to the closure problem, it is also of use in the case where no constraint is available. By hierarchically building sub-sequences using trifocal tensor registration, and then registering these sub-sequences together we can improve the speed and accuracy of the overall strategy. The m-view bundle adjustment problem is broken down into a number of smaller subproblems followed by a final m-view pass. The reason for the improvement is that the final pass generally converges much more quickly and to a better minimum if preceded by the hierarchical approach. To give some example costs: if the sequence is broken into 3 parts, each of $m/3$ views, the cost to process each part[1] is $m^2/9$, and the cost of processing all subparts is $m^2/3$. If

[1] Although bundle adjustment of many views is dominated by an m^3 cost, the small number of views used here means that the overall cost is approximately m^2

the final adjustment is made faster by a factor of more than $\frac{1}{3}$, the hierarchical strategy is faster. Table 2 shows typical results from our implementation: the costs are roughly offset, meaning that both approaches take approximately the same time, but the hierarchical approach achieves a lower reprojection error.

	Timings (sec)			RMS Errors	
Number of views	subseq	full	Total	Initial	Final
19	-	285	285	0.638	0.175
3 × 7	74	222	296	0.332	0.161

Table 2. Hierarchical versus monolithic processing. Comparison for one example sequence. The hierarchical approach is slightly slower on this short sequence, but is more accurate.

5 Discussion

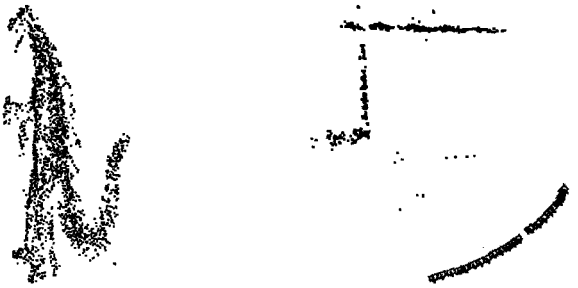
We have presented a system for structure and motion recovery which overcomes many of the problems with previous methods. The system builds on the maturity and robustness of the estimation algorithm for the trifocal tensor, and extends these algorithms to sequence matching in a similarly robust manner. In addition, the approach makes it easy to solve some other difficult problems, notably employing the extra information provided in a closed sequence.

Computational complexity is placed on the shoulders of the trifocal tensor computations, which are independent of each other. This independence means that it is easy to continue processing if one of the triplet computations fails. Also, the independence of these (very) large-grain processes means that parallel computation is immediately useful.

In summary the range of applicability of the approach is largely that of the traditional sequential systems, but it is markedly superior in terms of reliability, accuracy, ease of use, and possibly speed. Figures 7 and 8 show results of the system.

A Bundle Adjustment

A key component of the system described here is the ability to quickly compute maximum likelihood estimates of cameras and structure via bundle adjustment. The description of the process is relatively simple: For m views of n 3D points, we wish to estimate projection matrices $\{\hat{\mathrm{P}}_i\}_{i=1}^m$ and 3D points $\hat{\mathbf{X}}_j$ which project exactly to image points $\hat{\mathbf{x}}_{ij}$ as $\hat{\mathbf{x}}_{ij} = \hat{\mathrm{P}}_i\hat{\mathbf{X}}_j$. The projection matrices and 3D points which we seek are those that minimize the image distance between the



(a) (b)

Fig. 7. Results: 3D point structure for the dinosaur sequence (a). The castle sequence (b) is shown in plan view with the estimated cameras.

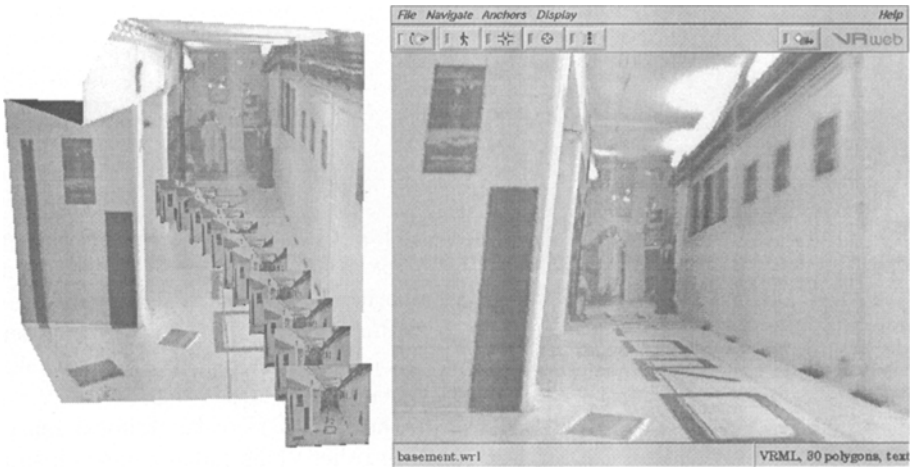


Fig. 8. Basement: Texture mapped planar model built from 11 views of the basement sequence. Left: VRML model of the scene with the cameras represented by their image planes. Right: a rendering of the scene from a novel viewpoint different from any in the sequence. The planar structure was built from 3D lines extracted by the algorithm of Schmid [19] using projection matrices computed by our algorithm.

reprojected point and detected (measured) image points $\mathbf{x}_{ij}$ for every view in which the 3D point appears, i.e.

$$\min_{\hat{\mathrm{P}}_i, \hat{\mathbf{X}}_j} \epsilon = \sum_{ij} d^2(\hat{\mathrm{P}}_i \hat{\mathbf{X}}_j, \mathbf{x}_{ij})$$

where $d(\mathbf{x}, \mathbf{y})$ is the Euclidean image distance between the homogeneous points $\mathbf{x}$ and $\mathbf{y}$. If the image error is Gaussian then bundle adjustment is the MLE. While simply expressed, the size of this optimization problem in a typical sequence of 30 images of 3000 points means that particular care must be taken to ensure

a computationally tractable solution. Following [6], efficient use is made of the block structure of the matrices involved, and the sparsity of the problem.

Acknowledgements

We are grateful for the castle sequence supplied by the University of Leuven, and the dinosaur sequence supplied by the University of Hannover. Financial support was supplied by EU ACTS Project VANGUARD.

References

1. N. Ayache. *Artificial vision for mobile robots.* MIT Press, Cambridge, 1991.
2. 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.
3. P. Beardsley, A. Zisserman, and D. W. Murray. Navigation using affine structure and motion. In *Proc. ECCV*, LNCS 800/801, pages 85–96. Springer-Verlag, 1994.
4. C. J. Harris. Determination of ego-motion from matched points. In *Alvey Vision Conf.*, pages 189–192, 1987.
5. C. J. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conf.*, pages 147–151, 1988.
6. R. I. Hartley. Euclidean reconstruction from uncalibrated views. In J. Mundy, A. Zisserman, and D. Forsyth, editors, *Applications of Invariance in Computer Vision*, LNCS 825, pages 237–256. Springer-Verlag, 1994.
7. R. I. Hartley. A linear method for reconstruction from lines and points. In *Proc. ICCV*, pages 882–887, 1995.
8. R. I. Hartley and P. Sturm. Triangulation. In *American Image Understanding Workshop*, pages 957–966, 1994.
9. A. Heyden and K. Åström. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In *Proc. CVPR*, 1997.
10. D. Jacobs. Linear fitting with missing data: Applications to structure from motion and to characterizing intensity images. In *Proc. CVPR*, pages 206–212, 1997.
11. S. Laveau. *Géométrie d'un système de N caméras. Théorie, estimation et applications.* PhD thesis, INRIA, 1996.
12. S. J. Maybank and A. Shashua. Ambiguity in reconstruction from images of six points. In *Proc. ICCV*, pages 703–708, 1998.
13. 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.
14. P. F. McLauchlan, I. D. Reid, and D. W. Murray. Recursive affine structure and motion from image sequences. In *Proc. ECCV*, volume 1, pages 217–224, May 1994.
15. R. Mohr, B. Boufama, and P. Brand. Accurate projective reconstruction. In J. Mundy, A. Zisserman, and D. Forsyth, editors, *Applications of Invariance in Computer Vision*, LNCS 825. Springer-Verlag, 1994.
16. M. Pollefeys, R. Koch, and L. Van Gool. Self calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *Proc. ICCV*, pages 90–96, 1998.

17. J. Porrill. Optimal combination and constraints for geometrical sensor data. *Intl. J. of Robotics Research*, 7(6):66–77, 1988.
18. I. D. Reid and D. W. Murray. Active tracking of foveated feature clusters using affine structure. *Intl. J. of Computer Vision*, 18(1):41–60, 1996.
19. C. Schmid and A. Zisserman. Automatic line matching across views. In *Proc. CVPR*, pages 666–671, 1997.
20. A. Shashua. Trilinearity in visual recognition by alignment. In *Proc. ECCV*, volume 1, pages 479–484, May 1994.
21. H. Y. Shum, M. Hebert, K. Ikeuchi, and R. Reddy. An integral approach to free-form object modeling. In *Proc. ICCV*, pages 870–875, 1995.
22. G. Sparr. Simultaneous reconstruction of scene structure and camera locations from uncalibrated image sequences. In *Proc. ICPR*, 1996.
23. M. E. Spetsakis and J. Aloimonos. Structure from motion using line correspondences. *Intl. J. of Computer Vision*, 4(3):171–183, 1990.
24. P. Sturm. *Vision 3D non calibrée: Contributions à la reconstruction projective et étude des mouvements critiques pour l'auto calibrage.* PhD thesis, INRIA Rhône-Alpes, 1997.
25. P. Sturm and W. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proc. ECCV*, pages 709–720, 1996.
26. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *Intl. J. of Computer Vision*, 9(2):137–154, 1992.
27. P. H. S. Torr, A. W. Fitzgibbon, and A. Zisserman. Maintaining multiple motion model hypotheses over many views to recover matching and structure. In *Proc. ICCV*, pages 485–491, January 1998.
28. P. H. S. Torr and D. W. Murray. Statistical detection of independent movement from a moving camera. *Image and Vision Computing*, 1(4):180–187, May 1993.
29. P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *Intl. J. of Computer Vision*, 24(3):271–300, 1997.
30. P. H. S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.
31. W. Triggs. Auto-calibration and the absolute quadric. In *Proc. CVPR*, pages 609–614, 1997.
32. Z. Zhang, R. Deriche, O. 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.
33. Z. Zhang and O. Faugeras. *3D Dynamic Scene Analysis.* Springer-Verlag, 1992.
34. A. Zisserman, P. Beardsley, and I. Reid. Metric calibration of a stereo rig. In *IEEE Workshop on Representation of Visual Scenes, Boston*, pages 93–100, 1995.