[Journal Name], [Volumn Number], 1–20 ([Volumn Year])

The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences

PHILIP H. S. TORR
Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA, philtorr@microsoft.com

ANDREW W. FITZGIBBON AND ANDREW ZISSERMAN
Robotics Research Group, Department of Engineering Science, Oxford University, OX1 3PJ., UK,
awf,az@robots.ox.ac.uk

;

Abstract. The aim of this work is the recovery of 3D structure and camera projection matrices for each frame of an uncalibrated image sequence. In order to achieve this, correspondences are required throughout the sequence. A significant and successful mechanism for automatically establishing these correspondences is by the use of geometric constraints arising from scene rigidity. However, problems arise with such geometry guided matching if general viewpoint and general structure are assumed whilst frames in the sequence and/or scene structure do not conform to these assumptions. Such cases are termed *degenerate*.

In this paper we describe two important cases of degeneracy and their effects on geometry guided matching. The cases are a *motion degeneracy* where the camera does not translate between frames, and a *structure degeneracy* where the viewed scene structure is planar. The effects include the loss of correspondences due to under or over fitting of geometric models estimated from image data, leading to the failure of the tracking method. These degeneracies are not a theoretical curiosity, but commonly occur in real sequences where models are statistically estimated from image points with measurement error.

We investigate two strategies for tackling such degeneracies: the first uses a statistical model selection test to identify when degeneracies occur; the second uses multiple motion models to overcome the degeneracies. The strategies are evaluated on real sequences varying in motion, scene type, and length from 13 to 120 frames.

1. Introduction

The goal of this work is to obtain 3D structure and camera projection matrices from an uncalibrated image sequence. The structure and cameras are used as a basis for building 3D graphical models from an image sequence; e.g. from sequences obtained from a hand-held camcorder where the motion is unlikely to be smooth, or known *a priori*. A typical example sequence and VRML model is shown in figures 12 and 14. The estimation of a camera for each frame and scene structure are also required in other applications, for example in order to insert virtual objects into real image sequences [24].

Several systems [4, 31, 38] have now been developed which are aimed at recovering 3D graphical models from image sequences. All are underpinned by the necessity to match tokens/features

(usually interest points) successfully through image sequences with a large number of frames. Indeed it transpires that the correspondence problem is one of the most difficult parts of structure recovery, and this is especially so when these correspondences must be maintained through many images. Many successful systems have been built which match features sequentially through sequences [2, 4, 5, 11, 23, 31, 33, 40]. Several of these systems require the calibration of the camera to be known *a priori*. In the work described here the camera can either be calibrated *or* uncalibrated.

A fundamental component in several of these tracking schemes is the use of epipolar geometry to simplify the search for correspondences between view pairs. This is because epipolar geometry and matches consistent with this geometry may be computed simultaneously, using only features in each view. Geometry guided matching can also be extended from two views to three or more, which is equivalent to using structure to guide matching. These geometry based matching methods are reviewed in Section 2.

However, there are certain commonly occurring situations in which geometry guided matching will fail. These situations are *degeneracies* of the estimation process. They include: a *motion* degeneracy, where the camera rotates abouts its centre, but does not translate. In this case epipolar geometry is not defined and the appropriate matching relation is an image to image homography; and, a *structure* degeneracy, where all the visible scene points are coplanar. In this case, which is a special case of a critical surface degeneracy [22], the image data do not contain enough information to recover the epipolar geometry.

An additional problem arises from the structure degeneracy: it is not possible to compute a 4×4 projective transformation between two sets of 3D points if the only correspondences available are coplanar; Consequently, in sequence tracking the structure and cameras before and after the degeneracy cannot be placed in a common 3D projective frame.

These are important theoretical and practical limitations which apply to any projective structure and motion scheme. If these degeneracies occur and are not handled appropriately the uncalibrated Structure and Motion (SAM) recovery algorithm will fail. The objective of this paper is to describe strategies for identifying when these degeneracies occur and for overcoming their consequences, so that the SAM recovery algorithm can successfully complete the entire sequence.

To this end the two types of degeneracies and their effects are described in more detail in Section 3. Sections 4 – 5 then discuss methods for identifying and handling the degeneracies, and section 6 methods for distinguishing between them. This discussion is illustrated by a number of sequences in which such degeneracies occur.

Notation and geometry

Points are represented by homogeneous 4-vectors $\mathbf{X}$ in 3-space, and by homogeneous 3-vectors $\mathbf{x}$ in the image. Matrices are denoted $\mathtt{M}$. The notation $[\mathbf{x}]_\times$ indicates the 3×3 skew matrix representing the vector product with $\mathbf{x}$, so that $[\mathbf{x}]_\times \mathbf{y} = \mathbf{x} \times \mathbf{y}$.

A point $\mathbf{X}$ is mapped to its image $\mathbf{x}$ by perspective projection, represented by a 3×4 camera matrix $\mathtt{P}$ as $\mathbf{x} = \mathtt{P}\mathbf{X}$. The fundamental matrix $\mathtt{F}$, and the homography between two images $\mathtt{H}$, are both 3×3 homogeneous matrices. Between homogeneous quantities, '=' indicates equality up to a non-zero scale factor.

2. Review: the correspondence problem over multiple views

Under rigid motion there are relationships between corresponding image points which depend only on the cameras and their motion relative to the scene, but not on the 3D structure of the scene. These relationships are used to guide matching. The relationships include the epipolar geometry between view pairs, represented by the fundamental matrix $\mathtt{F}$ [8, 17]; and the trifocal geometry between view triplets, represented by the trifocal tensor $\mathtt{T}$ [15, 29, 30]. These relationships, and image correspondences consistent with the relations, can be computed automatically from images, as described below.

Geometry guided matching, for view pairs and view triplets, is the basis for obtaining correspondences, camera projection matrices and 3D structure. The triplets may then be sewn together

to establish correspondences, projection matrices and structure for the entire sequence [4, 10].

Matching for view pairs. Correspondences are first determined between all consecutive pairs of frames as follows. An interest-point operator [12] extracts point features ("corners") from each frame of the sequence. Putative correspondences are generated between pairs of frames based on cross-correlation of interest point neighbourhoods within a search window. Matches are then established from this set of putative correspondences by simultaneously estimating epipolar geometry and matches consistent with this estimated geometry. The estimation algorithm is robust to mismatches and is described in detail in [34, 36, 39]. This basic level of tracking is termed the F-Based Tracker.

Matching for view triplets. Correspondences are then determined between all consecutive triplets of frames. The 3-view matches are drawn from the 2-view matches provided by the F-Based Tracker. Although a proportion of these 2-view matches are erroneous (outliers), many of these mismatches are removed during the simultaneous robust estimation of the trifocal tensor and consistent matches [35]. The trifocal geometry provides a more powerful disambiguation constraint than epipolar geometry because image position is completely determined in a third view, given a match in the other two views, whereas image position is only restricted to a line by the epipolar geometry between two views.

The output at this stage of matching consists of sets of overlapping image triplets. Each triplet has an associated trifocal tensor and 3-view point matches. The camera matrices for the 3-views may be instantiated from the trifocal tensor [16], and 3D points instantiated for each 3-view point match by minimizing reprojection error over the triplet (see appendix A.2).

Matching for sequences. Correspondences are extended over many frames by merging 3-view point matches for overlapping triplets [10, 21]. For example a correspondence which exists across the triplet 1-2-3 and also across the triplet 2-3-4 may be extended to the frames 1-2-3-4, since the pair 2-3 overlaps for the triplets. The camera matrices and 3D structure are then computed for the frames 1-2-3-4. This process is extended by merging neighbouring groups of frames until camera matrices and correspondences are established throughout the sequence. At any stage the available cameras and structure can be used to guide matching over any frame of the sequence. The initial estimate of 3D points and cameras for a sequence is refined by a bundle adjustment, as described in the appendix.

In summary, existing matches are verified using the trifocal tensor or projected structure. However, the basic mechanism for obtaining new matches is always the F-Based Tracker.

Performance issues. The robust nature of the estimation algorithms means that it is not necessary to restrict putative correspondences to nearest neighbours or even the highest cross-correlation match, as the rigidity constraint can be used to select the best match from a set of candidates. Typically the radius of the search window for candidate matches is 10–20% of the image size, which adequately covers image point motion for most sequences. It is necessary that the scene is for the most part rigid, but for moderate discrepancies—such as those caused by shadows and specularities, or other small moving objects—erroneous matches are excised automatically by the robust estimation. Given these restrictions the performance is excellent, with sequences of hundreds of frames being matched automatically.

3. The problem of degeneracy

The strategy outlined in the last section assumes *general* camera motion and *general* position for structure. The problem is that methods developed under generality assumptions fail when these assumptions do not hold. Situations of this type are termed degenerate. In the correspondence algorithm outlined above the F-Based Tracker may fail for an image sequence where either the camera motion is *not* general or the structure is *not* general.

The common non-general situations are of two types: the first (type A) is a motion degeneracy, the camera rotates about its centre and/or changes its internal parameters between views, but there is no translation; the second (type B) is

a structure degeneracy where all 3D points in the view are coplanar. The two cases are illustrated schematically in figure 1. In both these cases it is not possible to determine the epipolar geometry between consecutive frames: in type A the epipolar geometry is not defined, and in type B it cannot be uniquely determined from image correspondences alone. However, in both cases the image correspondences are related by a homography.

In case A, provided 3D-2D correspondences can be established between existing 3D structure and its image points in the current frame, then a camera matrix P can be computed from the 3D-2D correspondences. In case B the situation is more serious because even if 3D-2D correspondences can be established in the current frame, a (projective) camera matrix cannot be computed without additional information. This is because correspondences between coplanar 3D projective structure and their images only determine a homography with 8 degrees of freedom, so that 3 of the 11 degrees of freedom of the camera matrix are undetermined.

In the context of establishing projection matrices and structure sequentially, if there are frames for which all the viewed points arise from 3D points on a plane, then the 3D structure points recovered from views seen after the plane cannot be put into the same 3D projective coordinate frame as those seen before.

In the light of this it would appear that the ideal strategy would be to switch to homography, rather than epipolar, matching when a degeneracy is reached. In both cases such homography matching enables correspondences to be established through degenerate frame sets. For type A degeneracies maintaining correspondences is sufficient to completely alleviate the problem because camera matrices can then be computed for all frames and new correspondences established. Of course structure should not be instantiated from frames related by a rotation; necessitating the detection of this case. In case B a further strategy is required to place the 3D projective structure in a common frame; this possibility is returned to in section 5.

One can also consider the effects on *matching* of using the wrong image relation. Suppose a homography is used to compute correspondences in a sequence for which there is general motion and structure—the computation proceeds in a similar robust fashion to that described for estimating F in section 2 with both the homography H and image point correspondences being estimated simultaneously. The estimated homography can only fit a sub-part of the scene, often a dominant plane, and so potentially correct matches are lost. An example is shown in figure 3b for the Room sequence, where 50% of the matches are lost compared to using the correct matching relation. In terms of estimation, a model is being fitted which is too restrictive and so only part of the data fits well (an analogy is fitting a straight line to data lying on a plane).

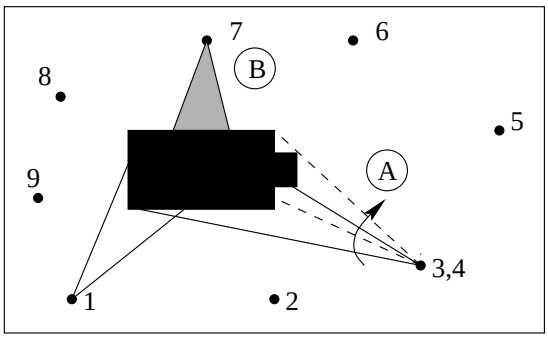


Fig. 1. An illustration of how the two types of degeneracies occur. A degeneracy of type A occurs between frames 3 and 4 where the camera rotates about its centre. A degeneracy of type B occurs at frame 7 where a plane fills the field of view.

Conversely, suppose epipolar geometry is used to compute correspondences for a type A or B degeneracy. Then the F model being fitted is underconstrained by the data: correspondences for these degeneracies are related by a homography, and satisfy a *family* of epipolar geometries with $\mathtt{F} = [\mathbf{e}']_\times \mathtt{H}$. The homography H between the views arises from the rotation for A, or is the homography induced by the world plane for B. This two-parameter family is parametrized by the epipole $\mathbf{e}'$ (with 2 real degrees of freedom) which is undetermined by correspondences related by the homography. The epipole is determined arbitrarily by the largest consistent sub-set of mismatches. Consequently, a result of overfitting is that points which might have been matched cor-

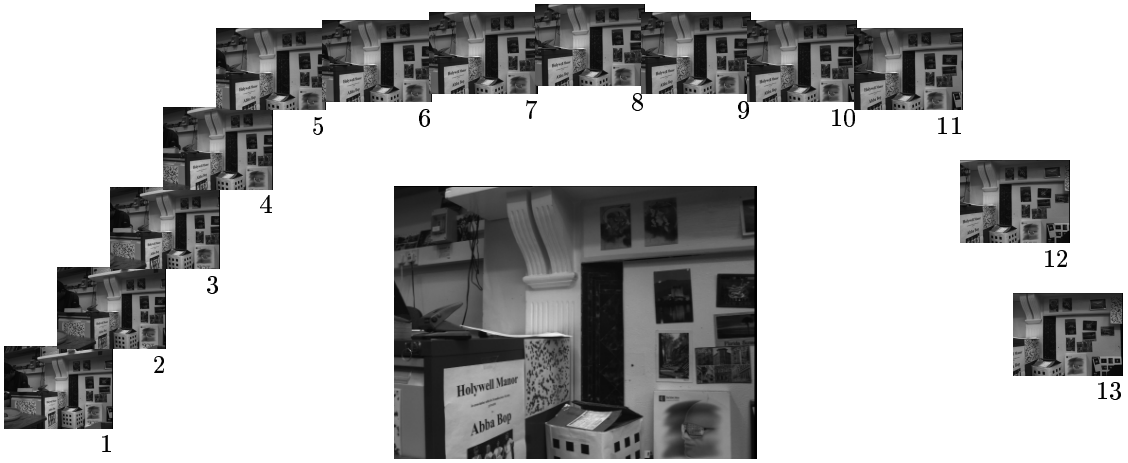


Fig. 2. Room sequence, including segments of pure rotation. Frames 5-10 are related to their successors by a rotation about the camera centre, whilst for the rest of the sequence there is a non-zero translation between frames.

rectly are mismatched and so, again, potentially correct matches are lost; and also additional mismatches are included between detected interest points which have no correct match under a homography. An example is shown in figure 8 for the rotation sequence of figure 6, where 30% of the matches are lost compared to using the correct matching relation (a homography). This discussion on over and under fitting is summarized in table 1.

The primary requirements then, which are the substance of the rest of this paper, are

(i) techniques to identify the correct motion model, specifically an alert mechanism that signals when `H`, rather than `F`, matching is necessary;
(ii) matching strategies that can survive type A degeneracies, for example by using the correct motion model between each frame;
(iii) matching strategies that can survive type B degeneracies, so that a new view can be integrated into an existing 3D projective frame;
(iv) methods to distinguish whether the homography arises from case A or B above.

These requirements are not just of theoretical interest: the relations are estimated from imperfect (noisy) image measurements so that a general motion sequence with small translations may be statistically indistinguishable from a type A degeneracy; similarly a scene with small relief relative to the distance to the camera may be statistically indistinguishable from a type B degeneracy. Especially for video sequences, where inter-frame disparities might be small, the best relation to use may not be that arising from general motion and structure. Ignoring degeneracies of either type means that the tracking system will often quite literally be "stopped in its tracks".

4. Type A Degeneracy

In this section we discuss possible strategies for overcoming a type A degeneracy, namely that at some point in the sequence the motion is not general as there is zero translation between views.

We will compare two approaches. The first (GRIC) is to attempt to determine the motion model, either `F` or `H`, based only on the information available for a pair of frames. The second method (MHT) maintains multiple motion hypotheses and does not attempt to determine which motion model is appropriate for a particular pair of frames.

Table 1. Possible combinations of matching relation and camera motion and structure.

		Actual Motion/Structure	
		General	Rotation or Plane
Relation	F	correct	over fitting
	H	under fitting	correct

4.1. The consequences of motion degeneracy

To illustrate the effects of matching using different motion relations, we will use the Room sequence, shown in figure 2, for which the ground truth motion type is known. The sequence was acquired under controlled conditions using a camera mounted on a mobile robot. There is general motion (non-zero translation) between frames 1 to 5; rotation about the camera centre between frames 5 to 11; and, a general motion between frames 11 to 13. The scene structure is general.

We will compare three scenarios for guiding matching: (a) using F throughout; (b) using H throughout; and finally (c) using the ground truth relation for each frame pair. Figure 3 shows the resulting tracks in each case, and table 2 the number of correspondences maintained though all the images.

The correctness of a track (through all the images) is determined automatically as follows. Since there is good overlap between the start and end of the sequence, a fundamental matrix can be computed using corresponding image points from the start and end of their tracks. This provides a necessary test for a correspondence since it must be inlying to this (robustly estimated) fundamental matrix. However, it does not test that the same point was tracked throughout the sequence. The correspondence throughout the sequence is tested by computing a 3D point from the start and end correspondence, and measuring reprojection error at every frame of the sequence. This is termed *structure consistency*. Note that structure consistency tests all frames of the sequence, both degenerate (pure rotational) and general motion.

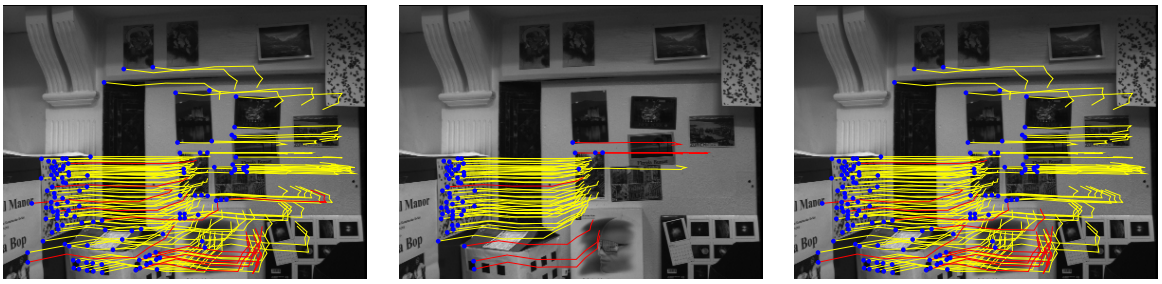
If F is used for the part of the sequence undergoing a rotational motion then this can lead to an increase in the number of false tracks. A clear example of this is shown in figure 8. If one is not alerted to the fact that this is a false F and the matches are used to instantiate new structure, the result will be many erroneous 3D points. The false matches arise because of overfitting as described in section 3.

Using H to guide matches throughout the sequence leads to fewer matches being extracted in the part of the sequence undergoing a general motion; as might be expected since the model underfits this part. However, as can be seen from table 2, when a loose threshold of 3 pixels is used (as opposed to a threshold of 1.25 pixels which is the two sigma window arising from interest point measurement noise) the homography is able to carry correct matches even through the non-rotational parts of the sequence. The explanation lies in the "plane plus parallax" model of image motion [19]: the estimated homography often behaves as if induced by a 'scene average' plane, or indeed is induced by a dominant scene plane; the homography map removes the effects of camera rotation and change in internal parameters, and is an exact map for points on the plane. The only residual image motion (which is *parallax* relative to the homography) arises from the scene relief relative to the plane. Often this parallax is less than the loose displacement threshold, so that all correspondences may still be obtained. Thus the homography provides strong disambiguation for matching and the parallax effects do not exceed the loose threshold. From here on only a threshold of 1.25 pixels is used.

An additional point to note is that tracking with the incorrect relation can be forgiving since incorrect tracks tend to disappear over long sequences, as they can rarely mimic rigid motion throughout, thus all of the relations yield few false

Table 2. Numbers of interest points tracked through the Room sequence under various motion models.
Full: Total number of tracks lasting through all frames of the sequence. **Consistent:** Number of full tracks inlying (1 pixel) to structure computation. **RMS:** RMS error after bundle adjustment.

Position inlier threshold for RANSAC = 1.25 pixels				
	Models	Full	Consistent	RMS
F	fffffffffffff	125	112	0.18
H	hhhhhhhhhhhh	52	47	0.17
Ground truth	ffffhhhhhfff	125	114	0.18
Position inlier threshold for RANSAC = 3 pixels				
	Models	Full	Consistent	RMS
F	fffffffffffff	127	110	0.18
H	hhhhhhhhhhhh	133	127	0.17
Ground truth	ffffhhhhhfff	129	111	0.18



(a) (b) (c)

Fig. 3. Correspondences using various motion models for the Room sequence (the position threshold is 1.25 pixels). (a) `F` only, 125 tracks, 112 consistent. (b) `H` only, 52 tracks, 47 consistent. (c) `F` and `H` as appropriate for ground truth motion, 125 tracks, 113 consistent.

matches. However, using the ground-truth relation does yield a slightly better result overall compared to `H` or `F` alone in this case.

We now describe and compare two methods for determining the best motion relation — best meaning that (1) tracking survives a motion degeneracy, and (2) the maximum number of veridical tracks are obtained.

4.2. Geometric robust information criterion

The Geometric Robust Information Criterion GRIC is a robust model selection criterion that is completely general. It is a scoring function for each model comprising two parts, one for the goodness of fit and one for the parsimony of the model. The first term is the minimum log likelihood of the data, the second is a penalty term, loosely proportional to a product of the number of parameters and the precision in those parameters. Its development is sketched in the appendix, with further details given in [32].

GRIC calculates a score function for each motion model taking into account the number n of inlier plus outlier correspondences, the residuals e_i, the standard deviation of the measurement error σ, the dimension of the data r (4 for two views) the number k of motion model parameters ($k = 7$ for `F`, $k = 8$ for `H`), and the dimension d of the structure ($d = 3$ for `F`, $d = 2$ for `H`)

$$\text{GRIC} = \sum \rho(e_i^2) + \lambda_1 dn + \lambda_2 k \quad . \qquad (1)$$

Where $\rho(e_i^2)$ is a robust function of the residuals:

$$\rho(e^2) = \min\left(\frac{e^2}{\sigma^2}, \lambda_3(r-d)\right) \quad . \qquad (2)$$

In all the evaluations here the parameters have values $\lambda_1 = \log 4$, $\lambda_2 = \log 4n$ and $\lambda_3 = 2$.

For each model GRIC is calculated and the model with the lowest score is indicated as most likely. Note that unlike other model selection criteria the GRIC criterion does not assert that the model with lowest score is the correct one, rather it provides the (negative) log of the posterior probability that the model is correct. Given two models under consideration the ratio of the exponentiated GRIC scores provides the relative odds of one being correct over the other.

4.3. Multiple hypothesis tracking

The primary idea behind MH tracking is to defer decisions until there is more information available to make them [3, 7, 27]. In this case for each pair of views, both models are fitted. Then for a three-view set, for example, the four possibilities are `FF`, `FH`, `HF` and `HH`. It might be thought that exploring all models would soon result in a computational explosion since 2^{n-1} possibilities must be explored for n frames. However, as is usual in MH tracking, a decision criterion is applied to prune out erroneous tracks. In this case after a set number of frames (13 here, i.e. the entire sequence) a trifocal tensor (equivalent to structure) is computed and any tracks which are inconsistent (measured

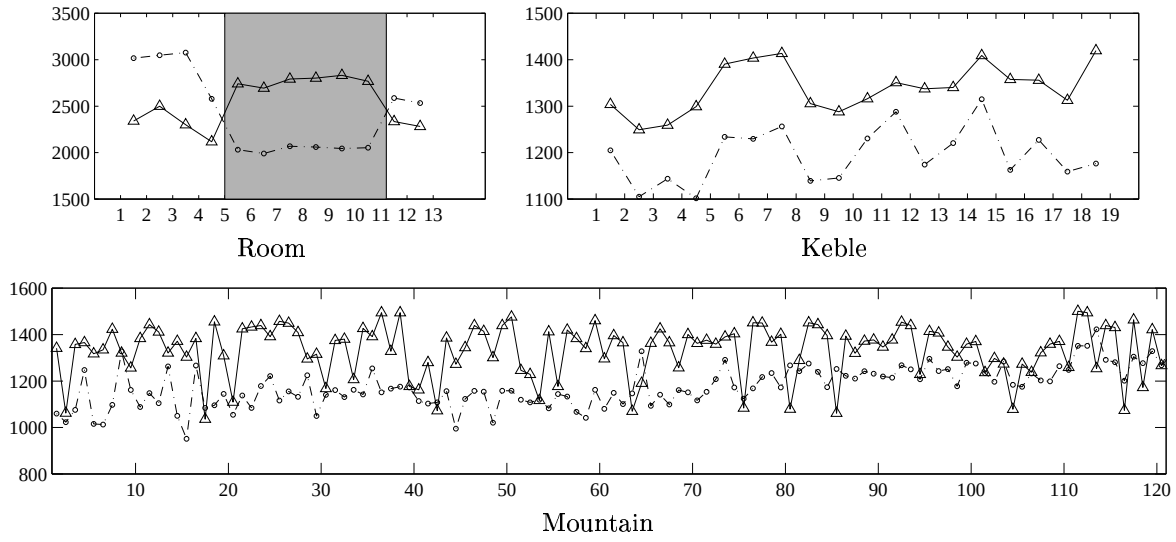


Fig. 4. GRIC scores for three example sequences. Dashed curves are `H` scores, black are `F`, both plotted against frame number. The Room sequence shows a switch to the homography model during the rotating frames (grey background). The Keble sequence is deemed to be a homography throughout, while the Mountain sequence (120 pairs of frames shown) is mostly considered to be a homography, occasionally switching to the `F` model when the inter-frame motion is large.

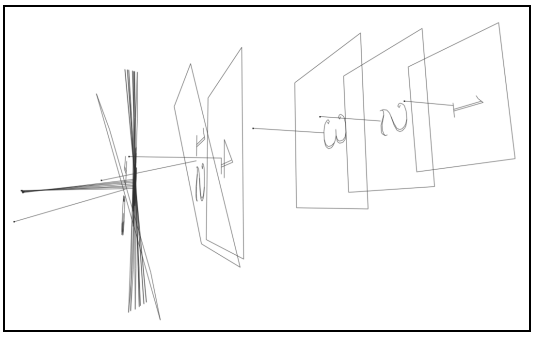


Fig. 5. Plan view of the camera positions computed after MH tracking of the Room sequence, showing the segment of rotation.

by reprojection error in each frame) are then excised. The measure of success of the scheme is the number of tracks which are structure consistent throughout the sequence.

One could then attempt to retrospectively ascertain which model was chosen, but this question cannot be satisfactorily answered because motion matches are not restricted to one model between frames. For example, between frames one and two (say) some points will have the same match under both models `F` and `H`, whilst for others the match will differ. When tracks are subsequently pruned, it may be that some of the matches from both `F` and `H` (which are not in common) survive.

There is another variation on the MH theme which is also implemented here. In the basic matching schemes reviewed in section 2 matches consistent with a motion model (`F` or `H`) are restricted to be unique, i.e. a point in frame one can only match one point in frame two. However, occasionally for `F` there are several suitable candidates for a match. A common example is where an image contains repeated structure (such as nearby windows on a building) so that points with similar intensity neighbourhood structure may lie on the same epipolar line. This situation does not arise in practice under the `H` model which is a point to point map. Consequently MH may be extended to include both multiple motion models *and* multiple consistent matches (for `F`). The latter scheme will be referred to as MHmulti, whereas the scheme where unique matches only are allowed will be referred to simply as MH.

Fig. 6. Pure rotation sequence. Three of 19 images of Keble college, Oxford, taken using a tripod-mounted camcorder.

Fig. 7. Small motion sequence. Three of 120 images of Arthur's seat, Edinburgh—acquired with a hand held camera whilst walking.

4.4. Evaluation of GRIC and MHT

Table 3 summarizes the results of these methods for the Room sequence, and figure 4 graphs the GRIC score. Camera positions estimated from the largest set of tracks are shown in figure 5. The correctness of a track is evaluated using structure consistency as described in section 4.1.

For this sequence GRIC selects the ground-truth model for every pair of frames. Also, the stability of the model selection may be observed in the graph of the difference $\mathrm{GRIC}_F - \mathrm{GRIC}_H$.

Table 3. Room sequence. Tracking results under the explored strategies.

Strategy	Full	Consistent	RMS
`F`	125	112	0.18
`H`	52	47	0.17
GRIC (Ground truth)	125	114	0.18
MH	143	125	0.17
MHmulti	360	127	0.17

This shows a sharp swing in favour of the homography model at pair $(5, 6)$, and then only small variations in the value until the reversion to the general model at pair $(11, 12)$.

As GRIC selects the ground truth motion model it has the same number of tracks as obtained by ground truth model matching. Both the MH schemes obtain *more* tracks than using the ground truth models. One would expect MH to obtain at least the same number of tracks as ground truth, because both `F` and `H` matches are included between every frame and so if the match is correct it will not be pruned. One reason that MH tracking achieves more matches is that it can take advantage of scene specialization. For example, suppose the scene consists of a planar and non-planar part and the motion is general. The `H` model will obtain many correct matches on the plane, and fewer off. The `F` model (which is correct for the motion) will obtain matches throughout the scene, but may well obtain fewer correct matches on the plane than the `H` model. When matches from both

models are combined there are more potentially correct matches available than using either model alone.

4.5. Evaluation on other sequences

We now compare GRIC and MHT on two further sequences. The first (Keble) is a pure rotation, i.e. every motion is a type A degeneracy. The second (mountain) is obtained by a hand held camera whilst walking, so should contain no degeneracies, but in practice camera translational motion between views is small compared to distance to the scene.

Keble Sequence. The sequence is shown in figure 6, and tracking results tabulated in table 5. The correct tracks cannot be evaluated using the structure consistency of section 4.1 (as applied to the Room sequence) for two reasons: first, tracks do not survive between the first and last frame because there is no image overlap; second, structure cannot be computed from views related by a rotation of the camera about its centre. Instead, the number of tracks which survive across n frames are reported (here $n = 8$), and correct tracks are determined by those which are matched between the first and nth frame and are also consistent in each of the intermediate frames (measured by reprojection error in each intermediate frame by `H` mapping the maximum likelihood image position estimated from the start and end frames).

As this is a pure rotation sequence the correct motion model is `H` throughout. For this degenerate sequence (only) the MHT employs a decision criterion based on whether matches are related by homography between suitably separated frames (here 8) as described above. The tracking performance is summarized in table 5. GRIC successfully selects the ground truth motion model (i.e. `H`) throughout the sequence. Using `F` as a motion model achieves fewer tracks than the ground truth model. The severe consequences of employing `F` as the motion model are demonstrated in figure 8 and table 4. For example, in the 8-view sub-sequence 13% of tracks are incorrect and 23% of the correct matches obtained by the `H` tracker are missed. Finally, both the MH schemes produce the correct tracks, although at greater computational expense.

Mountain Sequence. Figure 9 shows selected frames from a 120 frame sequence of a mountain (Arthur's seat in Edinburgh) acquired by a hand held camcorder. The scene is general (non-planar) and the motion results from walking with the camera, so is translational between most frames. The correct tracks are evaluated using structure consistency over $n = 20$ and $n = 40$ frames of the sequence. Figure 10 shows typical tracks, and tables 6 and 7 the number of tracks under the various matching strategies. Figure 9 shows track survival.

Over most frame pairs homography matching is preferred (as selected by GRIC) since parallax

Table 4. Keble rotation sequence: The consequences of overfitting. The number of errors caused by fitting `F` rather than `H` on the pure rotation sequence of figure 6. The results are averaged over all the 8 or 10-view contiguous subsequences of the full sequence.

Length	Wrong	Missed	Consistent
8	6.3	13.1	43.6
10	2.8	5.7	13.4

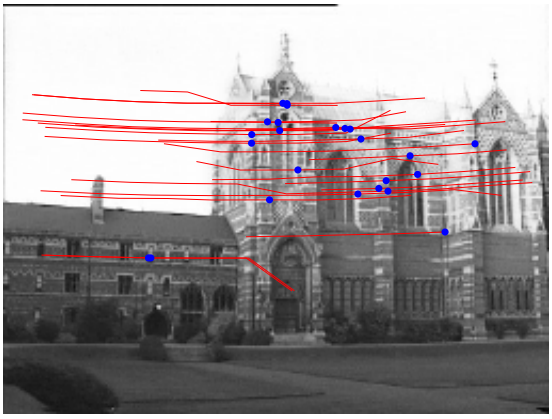


Fig. 8. A frame from the Keble sequence, showing some of the erroneous tracks resulting from overfitting, using `F` rather than `H`. The bends in the curve are caused when the track of one point is erroneously joined with the track of another point.

Table 5. Keble rotation sequence. Tracking results for the first 8 frames of the sequence. Consistency is measured by fitting a homography between the first and last views, and computing reprojection error in each intermediate view.

Strategy	Full	Consistent	RMS
`F`	67	61	0.209
`H` (Ground truth)	71	70	0.193
GRIC	71	70	0.193
MH	81	70	0.193
MHmulti	114	70	0.193

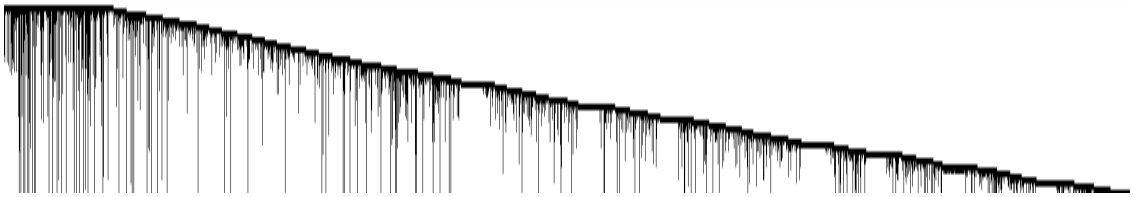


Fig. 9. Mountain sequence: track survival matrix for 60 frames. The Y axis is the frame number, X is the track number. Each vertical bar corresponds to a single tracked 3D point, and indicates the start and end frames in which that point was visible, the height of the bar therefore giving the length of the track.

effects are small. Occasionally, for example when a significant foreground object comes into view, or a particular inter-frame motion is large, `F` can be estimated without overfitting, and GRIC switches to `F`. Here, this results in an extra track being found which exceeds the threshold for `H` tracking. The `F` and MH tracks are supersets (i.e. contain additional correct tracks) of the basic set of correct tracks obtained by `H` and GRIC.

4.6. Summary

The three sequences and various matching strategies applied to them illustrate the following points:

Performance: the assessment criterion is the number of tracks which are structure consistent throughout every frame of their existence (homography consistent in the case of a pure rotation sequence).

If the `F` model is used in sub-sequences where there is pure rotational motion then erroneous tracks may result. These erroneous tracks may not survive through the entire sequence (and anyway can be pruned by structure consistency) but have the damaging effect of "claiming" points which otherwise might be correctly matched.

If the `H` model is used in general motion sequences with a tight threshold on displacement then tracks may be missed, but on the other hand few mismatches are generated because `H` defines a point to point map and consequently provides more disambiguating power than `F`. Of course, there is no problem at degenerate sub-sequences in using `H` as in this case it is the correct mo-

Table 6. Numbers of points tracked through the mountain sequence under various motion strategies, averaged over all 20-frame sub-sequences of the 120-frame sequence. About 200 correspondences are inlying to estimating `F` between the first and twentieth frames of each sub-sequence.
Full: Total number of tracks lasting through $n = 20$ frames of the sequence (before any structure consistent pruning in the MH case). **Consistent:** Number of full tracks inlying (1 pixel) to structure computation. **RMS:** RMS error after bundle adjustment.

Strategy	Full	Consistent	RMS
`F` (Ground truth)	76.5	65.5	0.165
`H`	65.1	62.8	0.145
GRIC	69.7	63.2	0.166
MH	79.0	65.9	0.166
MHmulti	214.5	66.9	0.168

Table 7. Mountain results, 40-frame sub-sequences. Key is as in Table 6.

Strategy	Full	Consistent	RMS
`F` (Ground truth)	37.9	34.0	0.136
`H`	35.4	33.9	0.138
GRIC	36.0	33.9	0.136
MH	38.8	34.0	0.139
MHmulti	121.0	34.7	0.137

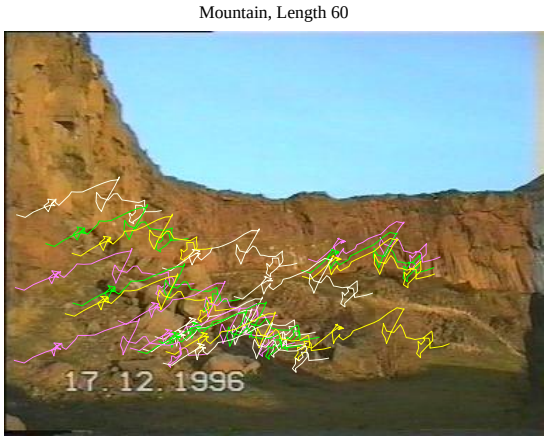
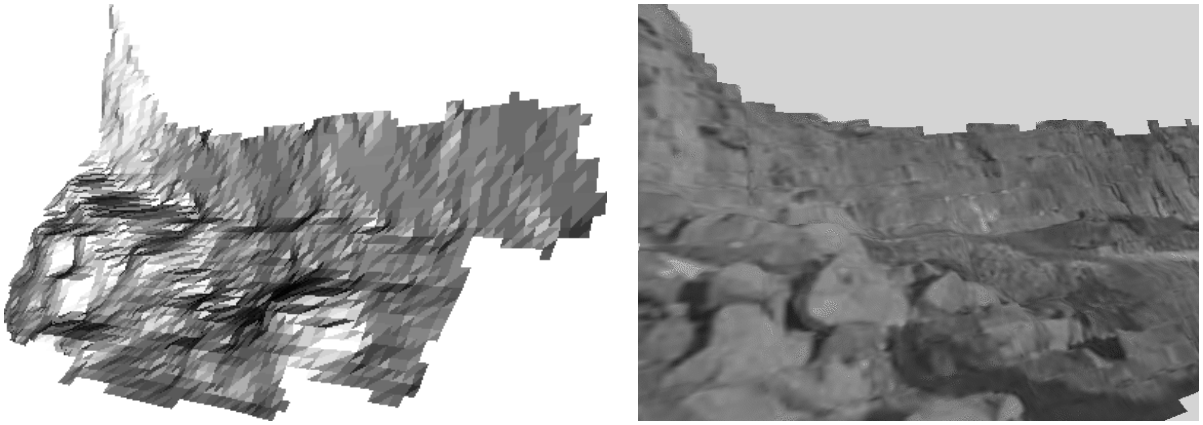


Fig. 10. Mountain sequence, 60-frame tracks superimposed on frame 0.



(a) (b)

Fig. 11. Mountain: VRML model. (a) Flat shaded. (b) Texture-mapped.

tion model. If the displacement threshold is loose, for example 3-5 pixels, then the disambiguation power is reduced. However, for a large class of scenes a homography plus loose threshold enables all correspondences to be determined under general motion. Such pairwise matching effectively enables a large baseline to be created through a sub-sequence, so that problems of overfitting with F are avoided.

GRIC is successful in indicating which motion model should be employed — if a motion model has a substantially lower score then it should be employed. If scores for each model are similar, then either model may be used.

The MH model enjoys the benefits of both the F and H models, but at a higher computational costs.

Ground truth: The use of the ground truth motion model is *not* necessarily the best means to achieve the goal of maximizing the number of structure consistent tracks. For example, as discussed above, H should be employed in general motion sequences when parallax effects are small, and MH tracking achieves at least as many tracks as ground truth, and in general more.

Table 8. Efficiency of MH schemes, expressed as the percentage of full tracks which were correct.

Scheme	Room	Keble	Mountain
MH	87	86	85
MHmulti	33	61	34

Efficiency and cost: GRIC is computationally cheap since it only involves a two frame comparison. A simple measure of the computational efficiency of each of the matching schemes is obtained from the ratio of consistent tracks to full tracks. Full tracks are those that survive throughout the sequence (or sub-sequence) and consistent tracks are in addition structure (or homography for pure rotation sequences) consistent in every frame. If the ratio is low, then much work is wasted in tracking points which will later be discarded. Ground truth tracking typically has an efficiency of 90 to 100%. The efficiency of the MH schemes is shown in table 8.

5. Type B Degeneracy

Consider a general motion sequence (i.e. non-zero translation between views) for a general scene (i.e. non-planar). At any particular frame, j say, the 3D position $\mathbf{X}_i$ of some points will be known, as will their images $\mathbf{x}_i^j$, and the camera matrix $\mathtt{P}^j$ for that frame, with the image and space points related as $\mathbf{x}_i^j = \mathtt{P}^j\mathbf{X}_i$. The addition of a new frame to the sequence requires that first image point correspondences $\mathbf{x}_i^j \leftrightarrow \mathbf{x}_i^{j+1}$ are established by estimating F, and once image correspondences are established the camera projection matrix $\mathtt{P}^{j+1}$ may be computed from the image and space point correspondences $\mathbf{x}_i^{j+1} \leftrightarrow \mathbf{X}_i$.

Suppose at some point in this general motion sequence the only correspondences available between the current frame and its successor arise from coplanar points. There are then two serious problems: first, the fundamental matrix cannot be estimated from image correspondences alone if their pre-images are all coplanar. This means that correspondences can not be established by estimating F (though they can be by estimating H); second, even if correspondences could be established, the camera projection matrix cannot be computed for the successor frame because a camera matrix cannot be estimated from the correspondence between space points on a plane and their image (remember that the camera is uncalibrated). There is a two parameter family for F as described in section 3. This family can be resolved by two point correspondences with pre-images off the plane. There is a three parameter family of P matrices, and in order to determine the camera matrix uniquely two points (actually 1.5) are required off the plane.

There are a number of strategies which can be applied to overcome this problem. The first is to note that if the camera is calibrated then the ambiguity (in both F and P computation) is reduced from a family to a discrete number of cases—two when F is estimated from a plane [9]. The camera can be calibrated by auto-calibration methods, for example from the camera matrices [25] or, if the plane itself is seen in five or more views, directly from the homographies induced by the plane [37].

A second strategy, which will be illustrated here, may be applied where the sequence is closed. In this case, the projective frame may be reinitialized by tracking from the first subsequent view in which correspondences are available both on and off the plane. In the new projective frame, the camera matrix for the plane-dominated view is well defined, and the problem becomes one of registering the two projective frames, which are related by a 3D-3D homography. Because the sequence is closed, this homography can be computed once the new projective frame overlaps the original one.

Model House: In Figure 12 six images of a model house are shown; this is taken from a 32 image sequence involving a complete circumnavigation of the house. The sequence exhibits a type B degeneracy problem as—half way round the sequence—a plane fills the view (Figure 13a) and the only tracked features lie on this plane. The projective structure for this half of the sequence (frames 0–12), which we shall call *projective frame PF1*, is shown in Figure 13b. In this case, correspondences both on and off the plane are available between frames 12 and 13, so a new projective frame (*PF2*) is instantiated beginning at frame 12, and continuing to the final frame (32). Now, as frame 32 and frame 0 are identical, we can compute the 3D homography which registers *PF1* (0–12) and *PF2* (12–32) as described in [10]. The result of this is shown in Figure 13c.

6. Distinguishing planes from rotations

We have seen that homographies arise in two situations: the camera has undergone a rotation about its centre (type A); or the 3D points for which there are correspondences are coplanar (type B). In either case the camera's internal parameters may have changed between views. In order to correctly recover structure and motion, it is important to be able to determine the type of degeneracy from which a particular homography-related pair of views has arisen. It might be thought that measuring the residuals to a 3D plane fit would distinguish types A and B, but this is impossible for a projective reconstruction as distance in $\mathcal{P}^3$ is not defined. Some other approaches, which do apply in the uncalibrated case, are now described.

GRIC method. It is possible to distinguish the two types of degeneracy if there exists precomputed structure (from previous images in the sequence) off the plane for which there are correspondences in the current view pair. The hypothesis that the structure is non-planar can be compared via GRIC to the hypothesis that the structure is planar. To evaluate each hypotheses an optimal estimate must be determined of each case. This involves a bundle adjustment as described in appendix A.2: for the non-planar hypotheses cameras and 3D structure are estimated; for the planar hypotheses homographies and 2D structure are estimated.

In order to compare the two models the GRIC scores for each are evaluated and the model with

Fig. 12. Three images from a sequence of 32 taken of a model house rotating on a turntable.

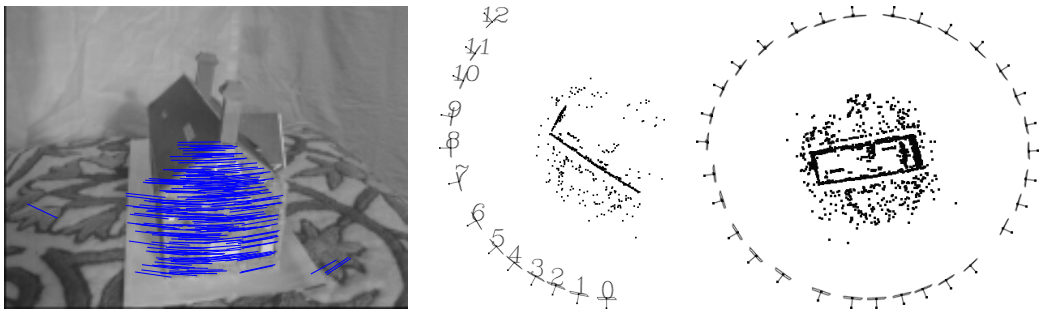


Fig. 13. (a) Tracked Matches to image 12, (b) Plan view of cameras and structure for frames 0-12 of the sequence. The cameras are numbered and represented by their image planes and principal axes. (c) Merged cameras and structure for 32 views.

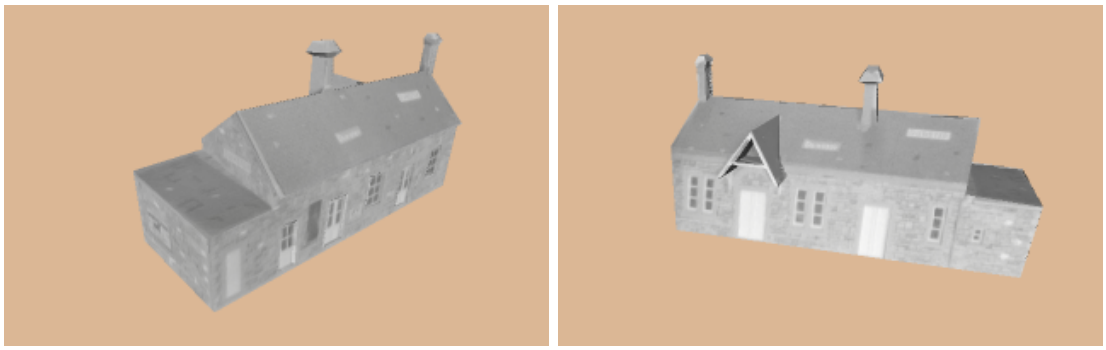


Fig. 14. Views of a texture mapped 3D model built from the structure automatically recovered in figure 13. The model is piecewise planar, with each plane texture mapped from the most fronto-parallel view based on the computed cameras.

lowest score is deemed more likely. For the homography over m views $d = 2$ and $k = 8(m-1)$. For the general motion $d = 3$ and $k = 7 + 11(m-2)$. For both models $\lambda_1 = \log 2m$ and $\lambda_2 = \log 2mn$.

Quick GRIC method. To avoid the onerous task of bundle adjustment a more computationally efficient single image test can be used, and has also been found to provide reasonable results. In case B, there exists a homography $\mathtt{M}$ between the (inhomogeneous) 3D structure and the 2D points:

$$\mathbf{x}_i^j = \mathtt{M}^j \begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix}. \qquad (3)$$

where $(X, Y, Z)^\top$ are the *in*homogeneous coordinates of the point $\mathbf{X}$. Suppose that the two images j and $j+1$ in the sequence are related by a homography $\mathtt{H}$ and that the structure is known. To determine which degeneracy pertains the procedure is as follows; First the 3D to 2D matrices $\mathtt{P}^j$, $\mathtt{P}^{j+1}$ and homographies $\mathtt{M}^j$, $\mathtt{M}^{j+1}$ are estimated from the structure into the jth and $j+1$th image by minimizing the reprojection error. Next the two GRIC scores (1) are evaluated. As the two images j and $j+1$ are related by a homography ($d=2$), the GRIC differ only in the term k representing the number of degrees of freedom in the parameters, which is $k=22$ for $\mathtt{P}^j$, $\mathtt{P}^{j+1}$ and $k=16$ for $\mathtt{M}^j$, $\mathtt{M}^{j+1}$, $\lambda_1 = \log 4$ and $\lambda_2 = \ln 4n$ for both models. Finally the model with lower GRIC score is deemed to be the one that holds. However it must be remembered that this test is only an approximation to the full bundle adjustment solution, and if the results are close it is necessary to resort to the bundle adjustment solution described in the previous paragraph. Results using this method on the Room sequence are shown in figure 15.

Eigenvalue method. The homography between images planes which arises from a pure rotation is a conjugate rotation [14]. This can be seen as follows: suppose the cameras before and after the pure rotation are $\mathbf{x} = \mathtt{K}[\mathtt{I}|\ \mathbf{0}]\mathbf{X}$ and $\mathbf{x}' = \mathtt{K}[\mathtt{R}|\ \mathbf{0}]\mathbf{X}$, where $\mathbf{x}$, $\mathbf{x}'$ are the corresponding image points for $\mathbf{X}$, then

$$\begin{aligned}\mathbf{x}' &= \mathtt{K}[\mathtt{R}|\ \mathbf{0}]\mathbf{X} \\ &= \mathtt{K}\mathtt{R}\mathtt{K}^{-1}\mathtt{K}[\mathtt{I}|\ \mathbf{0}]\mathbf{X} = \mathtt{K}\mathtt{R}\mathtt{K}^{-1}\mathbf{x}\end{aligned}$$

so that $\mathbf{x}' = \mathtt{H}\mathbf{x}$ with $\mathtt{H} = \mathtt{K}\mathtt{R}\mathtt{K}^{-1}$ which is a conjugate rotation. In this case the eigenvalues of $\mathtt{H}$ are $\{1, e^{i\theta}, e^{-i\theta}\}$ (up to scale), where θ is the angle of rotation of the camera.

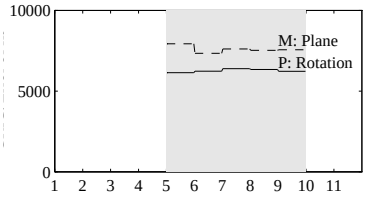


Fig. 15. GRIC selects the pure rotation over the planar model within the degenerate segment of the Room sequence.

It might be thought that this would be an effective test to distinguish homographies arising from coplanar structure from those arising from rotations. There are two reasons why testing the eigenvalues is not an effective test: first, under a rotation if the internal parameters change, as they might with automatic focus, the homography is no longer a conjugate rotation; second, the homography induced by a world plane can also have these eigenvalues—for example the homography induced in a camera undergoing planar motion (i.e. the translation is perpendicular to the rotation axis direction) by a plane perpendicular to the rotation axis. This is a common situation—a camera mounted on a car or robot observing the ground plane is an example.

7. Conclusions

We have identified two fundamental types of degeneracy that if not dealt with appropriately will result in failures of uncalibrated SAM algorithms. These are significant and universal problems—although details such as the particular number of tracks are peculiar to our matching method, the degeneracies apply to any matching method that assumes general motion and structure. Indeed, even in a calibrated system, a type A degeneracy will prevent tracking since epipolar geometry is undefined, and so must not be used to guide matching.

Several strategies have been suggested and implemented for type A (motion) degeneracy. The conclusion is that rotation degeneracies can be detected using a GRIC like scoring mechanism, and this test may then be used to signal that the $\mathtt{H}$ model should be used for matching. However, a multiple motion hypothesis tracking method also overcomes this degeneracy. The MH strategies have the advantage (over using either model alone) that more veridical tracks, and consequently better estimates for the cameras, are obtained, but at a higher computational cost. The recommendation therefore is to use MH if performance, but not cost, is premium.

For the type B (structure) degeneracy, we have adumbrated the problem and suggested some

means of detecting and dealing with it. Clearly, further strategies must be investigated in this case.

In conclusion, methods have been presented which expedite the recovery of 3D texture mapped models from uncalibrated image sequences. Initially 3D point correspondences are recovered as a means to determine the camera matrices for each frame. Once the camera matrices are recovered 3D models can be built by plane fitting, as in figure 14; or by intensity correlation [25], as in figure 11.

Acknowledgements

The mountain image sequence was kindly provided by the Edinburgh Virtual Environment Centre. We are also grateful to Reinhard Koch for providing the dense reconstruction software used to produce the 3D surface model of the mountain. Financial support was provided by EU ACTS Project Vanguard.

Appendix

MLE, Model Scoring and GRIC

Within this appendix a scoring criterion is set out for model comparison. The form of this scoring criterion involves a term for the likelihood, of each model together with a penalty term for the number of parameters, both derived in section A.1. To account for potential mismatches (outliers) the likelihood function has to be made robust as described in section A.2.1. Then the robust model score function GRIC is described in section A.3.3.

A.1. Maximum Likelihood Estimation

In order to calculate and compare all the two view relations an error function must be defined for each image relation. Although *ad hoc* error measures, such as the distance to epipolar lines (for F) or the distance between observed and projected point (for H), are often used, they have the disadvantage that it is difficult to compare them in a meaningful way. On the other hand, defining the log likelihood of each model has the advantage that its minimization puts the estimation of all the relations into a probabilistic framework which makes it easy to assess their relative likelihoods.

In this section the maximum likelihood formulation is given for computing any of the multiple view relations. In the following it is assumed that the noise on detected interest point positions is isotropic Gaussian with zero mean and uniform standard deviation σ. Thus given a *true* correspondence $\underline{\mathbf{x}}_i^1 \leftrightarrow \underline{\mathbf{x}}_i^2$ the probability density function of the noise perturbed *measured* correspondence $\mathbf{x}_i^1 \leftrightarrow \mathbf{x}_i^2$ is

$$\Pr(\mathbf{x}_i^1 \leftrightarrow \mathbf{x}_i^2 | \mathcal{R}) = \prod_i K e^{-\sum_{j=1,2} d^2(\underline{\mathbf{x}}_i^j, \mathbf{x}_i^j)/(2\sigma^2)} , \tag{A1}$$

where $d(\mathbf{x}, \mathbf{y})$ is the Euclidean image distance between two points $\mathbf{x}$ and $\mathbf{y}$, and $\underline{\mathbf{x}}^{2\top} \mathtt{F} \underline{\mathbf{x}}^1 = 0$ for the relation $\mathcal{R} = \mathtt{F}$, and $\underline{\mathbf{x}}^2 = \mathtt{H}\underline{\mathbf{x}}^1$ for the relation $\mathcal{R} = \mathtt{H}$, and K is a normalizing constant, independent of relation. The negative log likelihood of all the correspondences $\mathbf{x}_i^1 \leftrightarrow \mathbf{x}_i^2$, $i = 1..n$, where n is the number of correspondences is:

$$\begin{aligned} L &= -\sum_i \log(\Pr(\mathbf{x}_i^1 \leftrightarrow \mathbf{x}_i^2 | \mathcal{R})) \\ &= \frac{1}{\sigma^2} \sum_i \sum_{j=1,2} d^2(\underline{\mathbf{x}}_i^j, \mathbf{x}_i^j) \end{aligned} \tag{A2}$$

discounting the constant term.

Given two views with the true measurements $\underline{\mathbf{x}}_i^1, \underline{\mathbf{x}}_i^2$ unknown, "Maximum Likelihood Estimation" (MLE) equates to the task of finding the values of the relation $\mathcal{R}$ (F or H) and estimates $\hat{\mathbf{x}}_i^1 \leftrightarrow \hat{\mathbf{x}}_i^2$ of the true matches $\underline{\mathbf{x}}_i^1 \leftrightarrow \underline{\mathbf{x}}_i^2$, such that all $\hat{\mathbf{x}}_i^1 \leftrightarrow \hat{\mathbf{x}}_i^2$ exactly satisfy the relation $\mathcal{R}$, and minimise $\sum_{j=1,2} d^2(\hat{\mathbf{x}}_i^j, \mathbf{x}_i^j)$. The MLE error for the ith correspondence is

$$e_i^2 = \sum_{j=1,2} d^2(\hat{\mathbf{x}}_i^j, \mathbf{x}_i^j) \tag{A3}$$

Thus $\sum_i e_i^2$ provides the error function for the point data, and $\mathcal{R}$ for which $E = \sum_i e_i^2$ is a minimum is the maximum likelihood estimate of the relation (fundamental matrix or homography). Here $\hat{\mathbf{x}}_i^1 \leftrightarrow \hat{\mathbf{x}}_i^2$ has three degrees of freedom if the relation is a fundamental matrix (as it is the projection of a three dimensional point in the scene); and two degrees of freedom if the relation is a homography (as in this case it is the projection of a point on a plane). Thus the total number of pa-

rameters to be estimated is $7+3n$ for F and $8+2n$ for H. The MLE error is the same as the reprojection error of the optimally computed structure and can be solved for using bundle adjustment.

A.2. Bundle Adjustment

A bundle adjustment is an optimal estimate of the reconstruction. We wish to estimate projection matrices $\hat{\mathrm{P}}^j$ and 3D points $\hat{\mathbf{X}}_i$ such that the reprojection error to the measured image points $\mathbf{x}_i^j$ is minimized. This corresponds to minimizing the cost function

$$\min_{\hat{\mathrm{P}}^j, \hat{\mathbf{X}}_i} \sum_{ij} d^2(\hat{\mathrm{P}}^j \hat{\mathbf{X}}_i, \mathbf{x}_i^j) \qquad \text{(A4)}$$

where distances are included for every view in which the 3D point appears. If the measurement noise is Gaussian then as can be seen from the previous section bundle adjustment is the Maximum Likelihood estimate of the cameras and structure.

The cost function is minimized numerically using the Levenberg-Marquardt algorithm [26]. This is a large optimization problem — for example tracking through 30 images will typically involve 3000 3D points. More generally, the number of parameters that must be estimated is $11m + 3n$ for m views (11 for each projection matrix) and n 3D points. Following [13], efficient use is made of the block structure of the matrices involved, and the sparseness of the problem.

Bundle adjustment can also be applied to optimally estimate a homography relation over multiple views [6, 14]. This is required in the cases that the camera is rotating or the scene planar. We wish to estimate homography matrices for each view $\hat{\mathrm{H}}^j$ and 2D points $\hat{\mathbf{x}}_i$ such that the reprojection error to the measured image points $\mathbf{x}_i^j$ is minimized. This corresponds to minimizing the cost function

$$\min_{\hat{\mathrm{H}}^j, \hat{\mathbf{x}}_i} \sum_{ij} d^2(\hat{\mathrm{H}}^j \hat{\mathbf{x}}_i, \mathbf{x}_i^j) \qquad \text{(A5)}$$

where distances are included for every view in which there is a correspondence. In this case the number of parameters that must be estimated is $8m + 2n$ for m views (8 for each homography matrix) and n 2D points.

A.2.1. The Robust Cost Function

Rather than minimise (A3), the error actually minimised is a robust function [18]:

$$\rho(e) = \begin{cases} e^2 & e^2 < \lambda_3 d \\ \lambda_3 d & e^2 \geq \lambda_3 d \end{cases}, \qquad \text{(A6)}$$

where d is the number of degrees of freedom in the error (2 for H, 1 for F), and the parameter $\lambda_3 = 2^2$. The threshold 2 corresponds to the 95% confidence level. This means that an inlier will only be incorrectly rejected (a Type II error) 5% of the time.

This form of the function has several advantages. Firstly it provides a clear distinction between inliers and outliers. Secondly outliers to a given model are given a fixed cost, reflecting that they probably arise from a diffuse or uniform distribution, the log likelihood of which is a constant, whereas inliers conform to a Gaussian model.

A.3. Model Selection

One method of model comparison is to choose the model with maximum likelihood. The problem with this approach is that the more general model will always be accepted, hence the need for a more general method of inductive inference that takes into account the complexity of the model.

A.3.1. AIC/BIC *for Model Selection*

Akaike's information criterion is a useful statistic for model identification and evaluation. Akaike (1974) was perhaps the first to lay the foundations of information theoretic model evaluation. He developed a model selection procedure (for use in auto-regressive modelling of time series) that chose the model with minimum estimated expected residual, with respect to the model fitted, for future observations as the best fit. The procedure selects the model that minimizes expected error of new observations with the same distribution as the ones used for fitting. This error has the form

$$\mathrm{AIC} = L + 2k \quad , \qquad \text{(A7)}$$

where k is the number of parameters in the chosen model, and L is the negative log likelihood (A2).

It can be seen that AIC has two terms, the first corresponding to the badness of fit, the second a penalty on the complexity of the model. When there are several competing models, the parameters within the models are estimated by maximum likelihood and the AIC scores compared to find the model with the minimum value of AIC.

Schwarz [28] generalized AIC to BIC to take into account the number of degrees of freedom of the data (which are the correspondences in this case) in the penalty term as

$$\text{BIC} = L + 2k \ln n \quad . \tag{A8}$$

A.3.2. Geometric Information Criterion

Kanatani (1996) [20] generalises AIC to

$$\text{GIC} = L + 2(dn + k) \tag{A9}$$

where d is the dimension of the model. This can be understood by considering fitting a point or line to a set of 2D points. In this case $d = 1$ for the line model, and $d = 0$ for the point model, with $k = 2$ (the number of parameters of the model) in both cases.

Suppose points are generated from a fixed location with added mean zero, unit standard deviation, Gaussian noise in both the x and y coordinates. If a point and a line are fitted separately by minimizing the sum of squared Euclidean distances, the optimally fitted point will lie on the optimally fitted line. Let the sum of squared distances of the points to the line model be e_l^2 and the sum of squared distances of the points to the point model be e_p^2, then $e_p^2 = e_l^2 + e_k^2$, where e_k^2 is the 'parallel' sum of squared distances as shown for one point in Figure A.3.2.

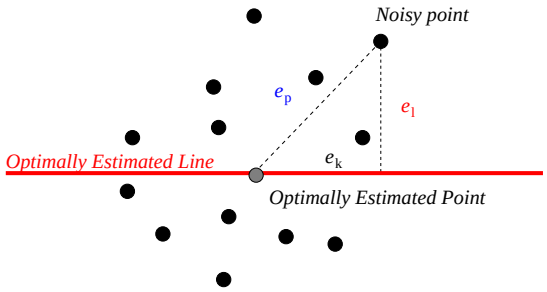


Fig. A.1. Showing the relationship between the noisy point, the optimally estimated line and the optimally estimated point in the Kanatani scheme.

It can be seen that unless the data all lie exactly on a point then e_l^2 is always less than e_p^2. The GIC for the line model compensates for this bias by the penalty term, which is twice the expectation of the 'parallel' sum of squares (e_k^2). If the model estimated is a line then the GIC has the form

$$\text{GIC(line)} = e_l^2 + 2(n + 2) \ , \tag{A10}$$

If the number of the data is large the degree of freedom of the model (i.e., the number of the parameters) has little effect because it is a simple constant. What matters is twice the dimension of the model, which is multiplied by the number of the data. The dimension equals the 'internal' degree of freedom of the data, which in turn equals the expectation of the 'parallel' (or in a direction on the manifold) sum of squares per datum. Returning to the example, the GIC for a point is

$$\text{GIC(point)} = e_l^2 + e_k^2 + 4 \ , \tag{A11}$$

thus a point is favoured if $e_k^2 \leq 2n$. Hence the algorithm is equivalent to a test of spread along the line.

A.3.3. Geometric Robust Information Criterion

GRIC combines Kanatani's extension (A9) with BIC, and generalizes in two ways: first different weightings are applied to the penalty terms; second, a robust cost function (A6) is used in place of the likelihood score L.

$$\text{GRIC} = L + \lambda_1 dn + \lambda_2 k \tag{A12}$$

In the case of H and F fitting, the weight of the penalty for the structure parameters is different to those of the motion parameters. This is because the motion parameters can be estimated with more certainty than the structure parameters. Loosely speaking, in two views each structure parameter can only be estimated from a match which contains four coordinates thus the weight for structure parameters is $\lambda_1 = \ln 4$, whereas for motion parameters these can be estimated from all the n correspondences (inliers and outliers) leading to a weighting of $\lambda_2 = \ln 4n$. Full details are given in [32].

References

1. H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, Vol. AC-19(6):716–723, 1974.
2. N. Ayache. *Artificial vision for mobile robots.* MIT Press, Cambridge, 1991.
3. Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association.* Academic Press, 1988.
4. P. Beardsley, P. Torr, and A. Zisserman. 3D model acquisition from extended image sequences. In *Proc. European Conference on Computer Vision*, LNCS 1064/1065, pages 683–695. Springer-Verlag, 1996.
5. P. Beardsley, A. Zisserman, and D. W. Murray. Navigation using affine structure and motion. In *Proc. European Conference on Computer Vision*, LNCS 800/801, pages 85–96. Springer-Verlag, 1994.
6. D. Capel and A. Zisserman. Automated mosaicing with super-resolution zoom. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Santa Barbara*, pages 885–891, June 1998.
7. I. J. Cox, J. M. Rehg, and S. Hingorani. A Bayesian multiple hypothesis approach to contour grouping. In *Proc. European Conference on Computer Vision*, LNCS 588, pages 72–77. Springer-Verlag, 1992.
8. O. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proc. European Conference on Computer Vision*, LNCS 588, pages 563–578. Springer-Verlag, 1992.
9. O. Faugeras. *Three-Dimensional Computer Vision: a Geometric Viewpoint.* MIT Press, 1993.
10. A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. European Conference on Computer Vision*, pages 311–326. Springer-Verlag, June 1998.
11. C. J. Harris. Determination of ego-motion from matched points. In *Third Alvey Vision Conference*, pages 189–192, 1987.
12. C. J. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference, Manchester*, pages 147–151, 1988.
13. 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.
14. R. I. Hartley. Self-calibration from multiple views with a rotating camera. In *Proc. European Conference on Computer Vision*, LNCS 800/801, pages 471–478. Springer-Verlag, 1994.
15. R. I. Hartley. A linear method for reconstruction from lines and points. In *Proc. International Conference on Computer Vision*, pages 882–887, 1995.
16. R. I. Hartley. Lines and points in three views and the trifocal tensor. *International Journal of Computer Vision*, 22(2):125–140, 1997.
17. R. I. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1992.
18. P. J. Huber. *Robust Statistics.* John Willey and Sons, 1981.
19. M. Irani and P. Anandan. Parallax geometry of pairs of points for 3d scene analysis. In B. Buxton and R. Cipolla, editors, *Proc. 4th European Conference on Computer Vision, LNCS 1064, Cambridge*, pages 17–30. Springer, 1996.
20. K. Kanatani. Automatic singularity test for motion analysis by an information criterion. In *Proc. 4th European Conference on Computer Vision, LNCS 1064, Cambridge*, pages 697–708, Springer–Verlag, 1996. Buxton, B. and Cipolla R.
21. S. Laveau. *Géométrie d'un système de N caméras. Théorie, estimation et applications.* PhD thesis, INRIA, 1996.
22. S. J. Maybank. *Theory of reconstruction from image motion.* Springer-Verlag, Berlin, 1993.
23. P. F. McLauchlan and D. W. Murray. A unifying framework for structure from motion recovery from image sequences. In *Proc. International Conference on Computer Vision*, pages 314–320, 1995.
24. P. Milgram, S. Shumin, D. Drascic, and J. Grodski. Applications of augmented reality for human-robot communication. In *International Conference on Intelligent Robots and Systems Proceedings, Yokohama, Japan*, pages 1467–1472, 1993.
25. M. Pollefeys, R. Koch, and L. Van Gool. Self calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *Proc. 6th International Conference on Computer Vision, Bombay, India*, pages 90–96, 1998.
26. W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C.* Cambridge University Press, 1988.
27. B. Rao. Data association methods for tracking systems. In A. Blake and A. Yuille, editors, *Active Vision*, pages 91–105. MIT Press, 1992.
28. G. Schwarz. Estimating dimension of a model. *Ann. Stat.*, 6:461–464, 1978.
29. A. Shashua. Trilinearity in visual recognition by alignment. In *Proc. 3rd European Conference on Computer Vision, Stockholm*, volume 1, pages 479–484, May 1994.
30. M. E. Spetsakis and J. Aloimonos. A multi-frame approach to visual motion perception. *International Journal of Computer Vision*, 16(3):245–255, 1991.
31. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision*, 9(2):137–154, November 1992.

32. P. H. S. Torr. An assessment of information criteria for motion model selection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico*, Jun 1997. To appear in CVIU.
33. 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. 6th International Conference on Computer Vision, Bombay, India*, pages 485–491, January 1998.
34. 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.
35. P. H. S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.
36. P. H. S. Torr and A. Zisserman. Robust computation and parameterization of multiple view relations. In *Proc. 6th International Conference on Computer Vision, Bombay, India*, pages 727–732, January 1998.
37. W. Triggs. Autocalibration from planar scenes. In *Proc. 5th European Conference on Computer Vision, Freiburg, Germany*, 1998.
38. C. Zeller. *Projective, Affine and Euclidean Calibration in Compute Vision and the Application of Three Dimensional Perception*. PhD thesis, RobotVis Group, INRIA Sophia-Antipolis, 1996.
39. 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.
40. Z. Zhang and O. Faugeras. *3D Dynamic Scene Analysis*. Springer-Verlag, 1992.