Multibody Structure and Motion: 3-D Reconstruction of Independently Moving Objects

Andrew W. Fitzgibbon and Andrew Zisserman
{awf,az}@robots.ox.ac.uk

Robotics Research Group, Department of Engineering Science,
University of Oxford, 19 Parks Road, Oxford OX1 3PJ, United Kingdom
http://www.robots.ox.ac.uk/~vgg

Abstract. This paper extends the recovery of structure and motion to image sequences with several independently moving objects. The motion, structure, and camera calibration are all *a-priori* unknown. The fundamental constraint that we introduce is that multiple motions must share the same camera parameters. Existing work on independent motions has not employed this constraint, and therefore has not gained over independent static-scene reconstructions.
We show how this constraint leads to several new results in structure and motion recovery, where Euclidean reconstruction becomes possible in the multibody case, when it was underconstrained for a static scene. We show how to combine motions of high-relief, low-relief and planar objects. Additionally we show that structure and motion can be recovered from just 4 points in the uncalibrated, fixed camera, case.
Experiments on real and synthetic imagery demonstrate the validity of the theory and the improvement in accuracy obtained using multibody analysis.

1 Introduction

This paper addresses the recovery of structure and motion from image sequences where the scene consists of several independently moving objects, and the motion, structure, and camera calibration are *a-priori* unknown. Figure 1 shows an example sequence. We show that by simultaneously analyzing multiple motions, more information can be recovered than is possible from each motion alone. The new constraint that we introduce is that although the motions are independent from frame to frame of the image sequence, the same camera views the 3D scene in each case. This means that in a Euclidean reconstruction the camera internal parameters are common across objects for each frame, and the absolute conic and the plane at infinity are common to the reconstructed objects. The shared camera parameters constrain the possible reconstructions of the moving scene.

The recovery of 3D from 2D up to now has depended on the presence of a single motion, often expressed as the "static-scene" assumption. The advantage that might be gained by multiple motions has been somewhat overlooked. Instead multiple motions have often been an irritant, and have at best been treated as a segmentation problem. In fact, multiple motions can contribute in a number of ways to solving the SFM problem. To give two examples:



Fig. 1. Frames from an image sequence containing independent motions. The cake tin is revolving as the camera moves through the scene. Simultaneous analysis of the motions improves reconstruction.

I. Two views: In the case of two views of a static scene it is well known that the Kruppa equations [8] provide two constraints on the camera internal parameters. If all the parameters apart from the focal length are known, then the focal lengths of the two views can be computed from the fundamental matrix [4, 12, 16]. Now suppose that in addition to the moving camera there is an object moving independently. Then both the object and the background scene each provide two Kruppa equations from their associated fundamental matrix. This means that there are four, in general, independent equations available on the internal parameters. Then, for instance, both the focal lengths and the principal point (assumed fixed but unknown) may be computed if the aspect ratio and skew are known.

II. Four views: Suppose that the background scene and object are tracked through four or more views so that camera projection matrices can be computed for each. Then (as will be demonstrated here) there are sufficient constraints to solve for the (scaled) Euclidean motion of the camera and object and background scene and internal parameters, *with the camera having different internal parameters in each view*.

Unlike previous approaches to auto-calibration, in the case of multiple motions with a sufficient number of views it is not necessary to make *any* assumptions about the internal parameters (e.g. that they are fixed) or camera motion – though of course such information may be included if available.

The advantages of multibody 3D reconstruction are most vividly demonstrated in cases such as those above, where static-scene reconstruction cannot give an unambiguous solution. However, even in situations where the static-scene formulation is applicable, we show that multibody reconstruction yields better results in the presence of measurement error. In the simplest instance, this is because more measurements contribute to the estimate of camera parameters, so that the estimation of the entire reconstruction is improved. In the large, however, we show too that when the motion estimate is poorly constrained by each of the individual objects, the composite may be better than any component, by virtue of the independent contributions of each moving object.

1.1 Background

Previous work on multiple motions has concentrated on the problem of motion segmentation. For example, independent motions can be segmented and clustered using robust estimation techniques [3, 25, 26], which successively search for the most consistent static scene correspondences. Throughout this paper, the motion segmentation

problem is assumed solved—and indeed motion segmentations herein have all been performed automatically using the standard principles exemplified in [23]. Where authors have extracted motion after segmentation [6, 10], the motions have been estimated independently, or have depended on calibrated cameras. Costeira and Kanade [5] do not assume prior segmentation, and provide a factorization method which recovers independent motions, and provides a motion segmentation, when the imaging conditions allow the weak perspective assumption. Again, however, the coupling provided by the shared camera constraint is not present, so no accuracy can be gained over independent static-scene reconstructions.

Avidan and Shashua have recently addressed a problem that they call "trajectory triangulation": the recovery of the position of a point moving on an unknown straight line [2] or an unknown conic [20], where the point is seen by a single camera. This is a more general problem than classical triangulation because the point is not observed simultaneously in the views. In their case the camera motion (matrix) is known *a-priori* and the point motion restricted. This makes the problem quite different to that studied here.

1.2 Camera models

Depending on the source of one's imagery, one of a number of alternative camera models will pertain. The models differ in the number of camera parameters which are known *a-priori*, the number that are constant through the image sequence, and the number that are varying through the sequence, and so must be estimated separately for every frame.

Calibrated: In the calibrated case, all camera parameters are known beforehand, and the task of 3D reconstruction is to determine the exterior orientation of the camera.

Fixed: In the model considered by much early research into uncalibrated structure and motion, the camera parameters are unknown, but assumed fixed throughout the sequence. In practice for video sequences from a cam-corder, aspect ratio and skew are very often fixed (and can be known *a-priori*).

Free: It is not uncommon for a real camera to vary three or more of its internal parameters in every view. For example, a "pan and scan" conversion from widescreen to video format will continually change principal point and occasionally aspect ratio, while focal length is being varied in the original frame. A collection of scanned photographs will typically have arbitrary principal points and focal lengths. Even for a camera which is ostensibly varying zoom only, mechanical misalignments will invariably induce a peripatetic principal point [30]. In these situations, static-scene 3D reconstruction cannot recover the internal parameters without scene knowledge.

Nonlinear lens distortion: In these models, and in this paper, we ignore the issue of correction for radial lens distortion, assuming that a linear projection approximation is sufficiently accurate to initialize a bundle adjustment which *does* include radial correction terms.

1.3 Notation

A standard pinhole camera model is employed, and points are represented in homogeneous coordinates. The 2D projection $\mathbf{x}$ of a 3D point $\mathbf{X}$ is given by

$$\mathbf{x} = \mathsf{K}[\mathsf{R} \mid \mathbf{t}]\mathbf{X} \;\; ; \qquad \mathsf{K} = \begin{bmatrix} f & s & u_0 \\ 0 & af & v_0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{cases} f \text{ is focal length;} \\ (u_0, v_0) \text{ is principal point;} \\ a \text{ is aspect ratio; } s \text{ is skew} \end{cases}$$

where for homogeneous quantities $=$ indicates collinearity of vectors, or equality up to a scale factor. The 3×3 rotation matrix R and translation $\mathbf{t}$ represent the position of the camera in space. Alternatively, we can consider the camera to be fixed in position, so that each object's position is specified by a rotation and translation with respect to the first view.

The image of the absolute conic (IAC) is given by $\boldsymbol{\omega} = (\mathsf{K}\mathsf{K}^\top)^{-1}$, and its dual (DIAC) is $\boldsymbol{\omega}^* = \mathsf{K}\mathsf{K}^\top$.

The image data is a sequence of V views of the scene containing B rigid bodies moving independently. In each view v, the camera parameters are given by K_v. The *projection equation* for point p of body β in view v is

$$ {}^{\beta}\mathbf{x}_{vp} \;=\; \mathsf{K}_v \left[{}^{\beta}\mathsf{R} \mid {}^{\beta}\mathbf{t} \right] {}^{\beta}\mathbf{X}_p \tag{1}$$

The objective of this work is to recover the structure and motion parameters which best model the observed image data $\{ {}^{\beta}\mathbf{x}_{vp} \}$.

2 Variations on Multiple Motions

In the following sections, we present a number of multiple-motion scenarios and show how the multibody analysis provides solutions which are not attainable in the static-scene formulation. The scenarios considered are chosen in order to illustrate the various types of advantage that multibody analysis confers. They are:

2.1 In the most general case, there are several independently moving bodies, and each body β contains sufficient points to enable a consistent set of camera matrices ${}^{\beta}\mathsf{P}_v$ to be computed throughout the sequence. This requires that at least 7 points are available (in general position) on each object for 2 views, and at least 6 points for 3 or more views.

2.2 At the other extreme are special cases of: scene structure, object motion, or imaging conditions. For example, the scene may consist of several independently moving planes, or the object might move under planar motion (fixed rotation axis direction and translation perpendicular to this direction), or the imaging conditions might be affine.

2.3 Between these general and specialized extremes are objects which offer only shallow relief, and therefore poorly constrain the camera matrix estimation. However, by simultaneously fitting to several motions, the spaces of degeneracy intersect, constraining correctly the final computed motion.

2.4 At the other end of the gamut from §2.1, we have the situation where very few points can be tracked on each object – insufficient to compute camera matrices, but the objects are not planar. We show that four or more points on a moving object constrain the camera calibration and Euclidean structure and motion recovery when tracked over multiple views.

These scenarios are now developed in detail.

2.1 Several high-relief bodies

In this scenario, we are tracking B independently moving bodies. We will start with the special case of two views. A fundamental matrix ${}^{\beta}\mathsf{F}$ is computed for each body β. From the general formula for the fundamental matrix [14]

$$\mathsf{F} = [\mathbf{e}_2]_\times \mathsf{K}_2 \mathsf{R} \mathsf{K}_1^{-1}$$

cancel $\mathsf{R}\mathsf{R}^\top = \mathtt{I}$ in $(\mathsf{F}\mathsf{K}_1)(\mathsf{F}\mathsf{K}_1)^\top$ to obtain the 2-camera Kruppa [29] equations

$$[\mathbf{e}_2]_\times \boldsymbol{\omega}_2^* [\mathbf{e}_2]_\times = \mathsf{F} \boldsymbol{\omega}_1^* \mathsf{F}^\top$$

where $\mathbf{e}_2$ is the epipole, given by $\mathbf{e}_2^\top \mathsf{F} = \mathbf{0}$. This homogeneous equation is linear in the elements of the DIACs $\boldsymbol{\omega}_1^*$ and $\boldsymbol{\omega}_2^*$, and generates two independent quadratic constraints relating the elements of $\boldsymbol{\omega}_1^*$ and $\boldsymbol{\omega}_2^*$ after removing the scale factor. Thus it provides two independent constraints on the 10 degrees of freedom (dof) of the two IACs. From just two views then, each independent motion provides 2 constraints on the two cameras. This leads to

Result 1 (Calibration from multiple objects: 5 objects, 2 views). *If two cameras view a scene containing five or more independently moving objects, the cameras for the two views can be completely calibrated.*

Multiple views: V $>$ 2 It is assumed that each independently moving body supports the computation of a sequence of P matrices. We will now extend the counting arguments and computation methods that have been developed in the recent auto-calibration literature [15, 17, 18, 22] for single object motion.

Counting argument. We choose to picture the camera as static, with objects moving relative to the camera. For each object the unknowns are the Euclidean motion up to an overall scale. Thus for V views there are $6\mathrm{V} - 7$ unknown parameters. However, in V views there are $11\mathrm{V} - 15$ constraints available from the P matrices. Thus the number of available constraints that may be applied to calibration (in principle) is $5\mathrm{V} - 8$. This results in 2 constraints from the fundamental matrix (the Kruppa equations), 7 from the trifocal tensor, and 12 constraints from the quadrifocal tensor.

Now suppose there are B objects. Since the objects are imaged by the same camera their calibration constraints (assumed independent) may be combined. Consequently, there are then $\mathrm{B}(5\mathrm{V} - 8)$ constraints available on a maximum of $5\mathrm{V}$ unknowns (5 calibration parameters per view). Thus provided $\mathrm{B}(5\mathrm{V} - 8) \geq 5\mathrm{V}$, the calibration can be determined in principle. This counting argument leads to the following result:

Result 2 (Calibration from multiple objects: 2 objects, 4 views). *Given four or more views of two moving objects the camera for each view can be completely calibrated.*

Algebra. Now we develop the multiple-view constraints algebraically. We let

$$ {}^{\beta}\mathsf{P}_i = [\, {}^{\beta}\mathsf{A}_i \mid {}^{\beta}\mathbf{a}_i] $$

denote the camera matrix of object β in the ith view. For the first view we assume that ${}^{\beta}\mathsf{P}_1 = [\mathtt{I} \mid \mathbf{0}]$ for all β. This fixes the reconstruction frame of each object to coincide with the frame of the first camera, an idea used by Sinclair [21]. For each object a Euclidean reconstruction (up to an overall scale) is then obtained from the projective reconstruction as

$$ {}^{\beta}\mathsf{P}_i \, {}^{\beta}\mathsf{H} = \mathsf{K}_i [\, {}^{\beta}\mathsf{R}_i \mid {}^{\beta}\mathbf{t}_i] \quad i = 2, .., \mathrm{V} \tag{2} $$

where ${}^{\beta}\mathsf{H}$ is the 4×4 homography

$$ {}^{\beta}\mathsf{H} = \begin{bmatrix} \mathsf{K}_1 & \mathbf{0} \\ {}^{\beta}\mathbf{v}^{\top} & 1 \end{bmatrix} \tag{3} $$

which must be determined. Note that K_i is common across objects, only the plane at infinity, specified by ${}^{\beta}\mathbf{v}$, need be determined independently for each object in the projective frame specified by $\{ {}^{\beta}\mathsf{P}_i \}$. Thus in order to determine the calibration matrix for each view (and the Euclidean motion for each object) it is only necessary to solve for $5 + 3\mathrm{V}$ unknowns; namely the 5 internal parameters of the first camera, K_1, and 3 parameters, ${}^{\beta}\mathbf{v}$, to specify the plane at infinity for each object (and thereby transform all the reconstructions to a common metric frame).

As is standard in the auto-calibration literature for a single motion, the rotation matrices can be eliminated to obtain equations relating only the unknowns. Applying the rectifying transformation H given by (3), the first 3×3 submatrix of the P matrix in (2) gives

$$ \mathsf{K}_i \, {}^{\beta}\mathsf{R}_i = {}^{\beta}\mathsf{A}_i \mathsf{K}_1 + {}^{\beta}\mathbf{a}_i \, {}^{\beta}\mathbf{v}^{\top} \quad i = 2, .., m \tag{4} $$

which may be rearranged as

$$ {}^{\beta}\mathsf{R}_i = (\mathsf{K}_i)^{-1} \left({}^{\beta}\mathsf{A}_i \mathsf{K}_1 + {}^{\beta}\mathbf{a}_i \mathbf{v}^{\top} \right) = (\mathsf{K}_i)^{-1} \left({}^{\beta}\mathsf{A}_i - {}^{\beta}\mathbf{a}_i \, {}^{\beta}\mathbf{p}^{\top} \right) \mathsf{K}_1 \quad i = 2, .., m \tag{5} $$

where ${}^{\beta}\mathbf{p} = -\mathsf{K}_1^{-\top} {}^{\beta}\mathbf{v}$. The rotation R_i may be eliminated on the right or the left, leaving

$$ \begin{aligned} \mathsf{K}_i \mathsf{K}_i^{\top} &= \left({}^{\beta}\mathsf{A}_i - {}^{\beta}\mathbf{a}_i \, {}^{\beta}\mathbf{p}^{\top} \right) \mathsf{K}_1 \mathsf{K}_1^{\top} \left({}^{\beta}\mathsf{A}_i^{\top} - {}^{\beta}\mathbf{p} \, {}^{\beta}\mathbf{a}_i^{\top} \right) \qquad \text{(DIAC)} \\ \left(\mathsf{K}_i \mathsf{K}_i^{\top} \right)^{-1} &= \left({}^{\beta}\mathsf{A}_i - {}^{\beta}\mathbf{a}_i \, {}^{\beta}\mathbf{p}^{\top} \right)^{-\top} \left(\mathsf{K}_1 \mathsf{K}_1^{\top} \right)^{-1} \left({}^{\beta}\mathsf{A}_i - {}^{\beta}\mathbf{a}_i \, {}^{\beta}\mathbf{p}^{\top} \right)^{-1} \text{(IAC)} \end{aligned} \tag{6} $$

which involve only the unknowns K_i, ${}^{\beta}\mathbf{p}$, and the known elements from the projective reconstruction ${}^{\beta}\mathsf{A}_i$, ${}^{\beta}\mathbf{a}_i$. The first equation (DIAC) is alternatively obtained by multiplying out the projection of the absolute (dual) quadric [27].

It is at this point that we break with the conventional route of auto-calibration, because in the multi-body case new equations can be formed. In particular, because K_i is common to all objects, each additional object generates five independent equations of the form (e.g. from the DIAC equation)

$$\boldsymbol{\omega}_i^* = \left({}^{\alpha}\mathsf{A}_i - {}^{\alpha}\mathbf{a}_i\,{}^{\alpha}\mathbf{p}^\top\right)\boldsymbol{\omega}_1^*\left({}^{\alpha}\mathsf{A}_i^\top - {}^{\alpha}\mathbf{p}\,{}^{\alpha}\mathbf{a}_i^\top\right) = \left({}^{\beta}\mathsf{A}_i - {}^{\beta}\mathbf{a}_i\,{}^{\beta}\mathbf{p}^\top\right)\boldsymbol{\omega}_1^*\left({}^{\beta}\mathsf{A}_i^\top - {}^{\beta}\mathbf{p}\,{}^{\beta}\mathbf{a}_i^\top\right) \tag{7}$$

for two objects α and β. In the case of two objects in V views there are then $5(\mathrm{V}-1)$ independent equations on 11 unknowns, so the problem may be solved provided $5\mathrm{V} - 16 \geq 0$, consistent with the counting argument above. Or, stated another way, each object provides 5 independent constraints for each view apart from the first, at a "start up cost" of only 3 parameters per object (namely ${}^{\beta}\mathbf{p}$). Alternatively, equations may be obtained involving only the unknowns ${}^{\beta}\mathbf{p}^\top$. These are the analogue of the "Modulus Constraint" [19].

If additional information is available, e.g. a fixed or skew zero camera, then additional equations are generated from (6), in the same manner as conventional single body auto-calibration.

2.2 Special cases of scenes and camera motions

Multiple independently moving scene planes. It is assumed that each plane β induces a homography ${}^{\beta}\mathsf{H}$ between each view. Triggs [28] has shown that for a single moving plane the homographies impose two constraints per view, after a "start up cost" of four has been paid. Geometrically the two constraints per view arise from the imaged circular points of the plane which lie on the IAC. These are transferred between views by H. The start up cost arises becauses initially the imaged circular points are unknown, and their four degrees of freedom must be specified.

Consequently, a fixed camera may be calibrated once $\mathrm{V} \geq (5+4)/2$, i.e. 5 views. Now consider a camera in which two (or more) internal parameters vary between views. If two parameters vary, then the number of unknowns is $2\mathrm{V}+7$ (the 7 results from three fixed camera parameters and the start up cost of 4 for the circular points). There are a total of $2\mathrm{V}$ constraints, but this is never sufficient to determine the $2\mathrm{V}+7$ unknowns. In the happy circumstance that 2 independently moving scene planes are available the accounting changes to four constraints per view plus a start up cost of eight, and the number of varying parameters may be increased to three (since the number of unknowns is $3\mathrm{V}$ for the varying parameters, and 2 for the fixed parameters and together with the fixed cost this totals $10+3\mathrm{V}$, while the number of constraints is $4\mathrm{V}$). We have then for example,

Result 3 (Calibration from multiple moving planes). *Given two or more independently moving planes, a camera with varying zoom and principal point may be calibrated from image data alone. A minimum of 10 views are required.*

If the camera can be calibrated then the motion is also determined (up to at most a finite ambiguity). If there are 3 planes, then all the internal parameters may vary provided there are at least 12 views (in general we require $2\mathrm{BV} \geq 5\mathrm{V} + 4\mathrm{B}$ for B planes).

Multiple Planar motions. It is assumed that each object is undergoing planar motion on a common plane, and that each object supports the computation of a consistent set of camera matrices. A common example would be a camera observing several cars at a road junction. In the case of a single object and fixed calibration Armstrong *et al.* [1] showed that from 3 or more views the plane at infinity is uniquely determined, and the internal parameters are determined up to a one parameter family.

If a scene contains two or more such degenerate motions, then each provides constraints on the plane at infinity and camera calibration. However, the one parameter ambiguity is not removed unless constraints are provided by an object not undergoing planar motion. This quickly gives us several results of the form

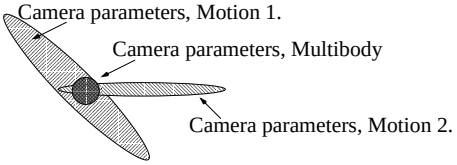
Result 4 (Disambiguation of planar motion). *A camera undergoing planar motion can be uniquely calibrated if another moving object (which may be a special structure such as a plane) is viewed undergoing independent, but general, motion.*

2.3 Low-relief objects

Often objects have low surface relief compared to their distance from the camera centre. If cameras are computed from such objects then poor estimates of the P matrices are obtained.

The general situation then is that each independent motion produces a probability density function over the space of possible 3D reconstructions. As shown in section §3.2, each body β is represented by a set of camera and scene parameters $\boldsymbol{\theta}_\beta$ which maximizes a likelihood function $P(\boldsymbol{\theta}_\beta)$.

Typically these pdfs will not have a sharply-defined mode, but will instead define a space of nearly equivalent reconstructions. A common example occurs when the camera motion is towards a low-relief scene, and the camera focal length is unknown. Conflation of focal length and distance to the scene means that focal length may vary over a wide range and still give a reconstruction which explains the image data. However, when there are multiple motions, their ambiguities will not in general coincide, so that combining the estimates will reduce the ambiguity. Figure 2, illustrates this process, and section 3.3 describes the implementation.



Each motion produces a poorly constrained reconstruction, meaning that *camera* parameters can be varied over a wide range (stroked ellipses) while generating small 2D reprojection errors. However, because the ranges are different for each independent motion, the multibody estimate (hatched ellipse) "intersects" the covariances to give an accurate result.

Fig. 2. Combining pdfs of multiple independent motions under the constraint that they share the same camera.

2.4 Few points

The final scenario explored here concerns the case when few points are available on each moving object. Suppose we can observe B bodies, each of P points, in V views. The measurements are the (x, y) position of the projection of each point in each view, so the number of measurements is $2\mathrm{VBP}$. The unknowns consist of 3 degrees of freedom for each 3D point (to represent its $(\mathrm{X}, \mathrm{Y}, \mathrm{Z})$ in the camera's coordinate frame), and 6 dof per view for each object to parametrize its Euclidean motion up to an unknown similarity transformation, giving $3\mathrm{BP} + \mathrm{B}(6\mathrm{V} - 7)$ dof in total. Consequently there are $2\mathrm{VBP}$ constraints on $3\mathrm{BP} + \mathrm{B}(6\mathrm{V} - 7)$ unknowns, and provided

$$2\mathrm{VBP} \;\geq\; \underbrace{\mathrm{B}\,(3\mathrm{P} + 6\mathrm{V} - 7)}_{\text{Objects + Motions}} \tag{8}$$

then there are constraints available for calibration. For 3 points per body, this reduces to $6\mathrm{BV} \geq \mathrm{B}(6\mathrm{V} + 2)$ so that adding views or bodies just provides as many measurements as degrees of freedom. However, for four points we obtain $2\mathrm{BV} \geq 6\mathrm{B}$ or $2\mathrm{B}(\mathrm{V} - 3)$ constraints on the camera parameters. In particular if $\mathrm{B} = 1$, then there are $2\mathrm{V} - 5$ constraints available for camera calibration. This means that after a "start up cost" of five, there are two constraints per view available for calibration. Thus a fixed camera can be determined after 5 views, and a camera with one parameter varying determined once $2\mathrm{V} - 5 > 4 + \mathrm{V}$, i.e. nine views. Thus we have

Result 5 (Structure and motion from four points). *Given four unknown rigidly coupled points undergoing Euclidean motion, observed by an unknown camera; it is possible to recover their structure, motion, and the camera calibration (even in the presence of a single varying intrinsic parameter), from image data alone.*

This result also applies in the case of four coplanar points since (as in section 2.2) structure and motion can be recovered from four coplanar points.

3 Implementation

The previous sections have provided the theoretical background to multibody reconstruction, without reference to implementation. In this section we briefly discuss these implementation issues, and in particular the cost function for bundle adjustment in the case of multibody motion, and the resulting covariances.

3.1 Point detection, tracking and segmentation

For this work, we employ standard point detection and tracking procedures, for example as described in [24]: interest points are detected using the Harris operator [11], and matched pairwise using a RANSAC procedure to compute the fundamental matrix, followed by triplet matching using a RANSAC procedure to compute the trifocal tensor. Suppose the sequence consists of a moving camera viewing a background scene and an independently moving foreground object; then the RANSAC scoring mechanism

will first select the dominant motion (say this corresponds to the background motion). Merging pairwise and triplet tracks gives multi-view 2D tracks for each 3D point of the background. Tracks corresponding to this dominant motion are then removed, and the procedure repeated to track the foreground object. This track removal and re-RANSAC loop can be repeated until too few matches remain to reliably compute the between-view relations.

3.2 Bundle Adjustment

A component of the implementation 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 V views of P 3D points, we wish to estimate camera calibrations K_v, Euclidean transformations $[\mathsf{R}_v \mid \mathbf{t}_v]$ and 3D points $\mathbf{X}_p$ which project exactly to image points $\hat{\mathbf{x}}_{vp} = \mathsf{P}_v \mathbf{X}_p = \mathsf{K}_v[\mathsf{R} \mid \mathbf{t}]\mathbf{X}_p$. The projection matrices and 3D points which we seek are those that minimize the image distance between the reprojected point and detected (measured) image points $\mathbf{x}_{vp}$ for every view in which the 3D point appears, i.e.

$$\min_{\mathsf{K}_v, \mathsf{R}_v, \mathbf{t}_v, \mathbf{X}_p} \epsilon = \sum_{v \in \text{VIEWS}} \sum_{p \in \text{POINT INDICES}} d^2(\mathbf{x}_{vp}, \quad \mathsf{K}_v[\mathsf{R}_v \mid \mathbf{t}_v]\, \mathbf{X}_p) \tag{9}$$

where $d(\mathbf{x}, \mathbf{y})$ is the Euclidean image distance between the homogeneous 2D points $\mathbf{x}$ and $\mathbf{y}$. If the image error is Gaussian then bundle adjustment is the MLE. This error may be efficiently minimized following [13].

For multiple objects, the bookkeeping is a little more complex, but the principle is analogous. The camera matrices are common between views, but each object has a separate set of Euclidean transformations. Thus the error function for two objects has the form

$$\begin{aligned} \epsilon &= \epsilon(\mathsf{K}_{1..\text{V}}, {}^{\alpha}\mathsf{R}_{1..\text{V}}, {}^{\alpha}\mathbf{t}_{1..\text{V}}, {}^{\beta}\mathsf{R}_{1..\text{V}}, {}^{\beta}\mathbf{t}_{1..\text{V}}, \mathbf{X}_{1..\text{P}}) \\ &= \sum_{v \in \text{VIEWS}} \left(\sum_{p \in \text{POINT INDICES } \alpha} d^2(\mathbf{x}_{vp}, \quad \mathsf{K}_v[\, {}^{\alpha}\mathsf{R}_v \mid {}^{\alpha}\mathbf{t}_v]\, \mathbf{X}_p) + \sum_{p \in \text{POINT INDICES } \beta} d^2(\mathbf{x}_{vp}, \quad \mathsf{K}_v[\, {}^{\beta}\mathsf{R}_v \mid {}^{\beta}\mathbf{t}_v]\, \mathbf{X}_p) \right) \end{aligned} \tag{10}$$

3.3 Covariance intersection

Where the motions in the scene provide reasonably well constrained camera parameters, we wish to combine the parameter estimates in a way that makes use of the uncertainty reduction illustrated in Figure 2. If the camera parameters are so constrained, an approximate covariance matrix of each set of parameters at the static-scene estimate may be computed from the Jacobian of the reprojection error (9), and will provide a reasonable model of the probability distribution $P(\boldsymbol{\theta})$ of the possible camera parameters.

From bundle adjustment, then, we compute the covariance estimate $\boldsymbol{\Lambda}_\beta$ at the minimum $\hat{\boldsymbol{\theta}}_\beta$ for each body β, so that

$$-\log {}^{\beta}P(\boldsymbol{\theta}) \approx (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\beta)^\top \boldsymbol{\Lambda}_\beta^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\beta)$$

Then a combined estimate of the camera parameters is obtained [7]. For example, given two objects with camera parameters $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$, and covariances $\boldsymbol{\Lambda}_A$ and $\boldsymbol{\Lambda}_B$, the combined parameters are given by

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\Lambda}_B(\boldsymbol{\Lambda}_A + \boldsymbol{\Lambda}_B)^{-1}\hat{\boldsymbol{\theta}}_A + \boldsymbol{\Lambda}_A(\boldsymbol{\Lambda}_A + \boldsymbol{\Lambda}_B)^{-1}\hat{\boldsymbol{\theta}}_B$$

In the multibody bundle adjustment this covariance intersection is achieved automatically.

4 Experiments

We now test the theory on several synthetic and real examples. The experiments are designed to verify the theoretical results, and to determine the level of improvement provided over static-scene techniques.

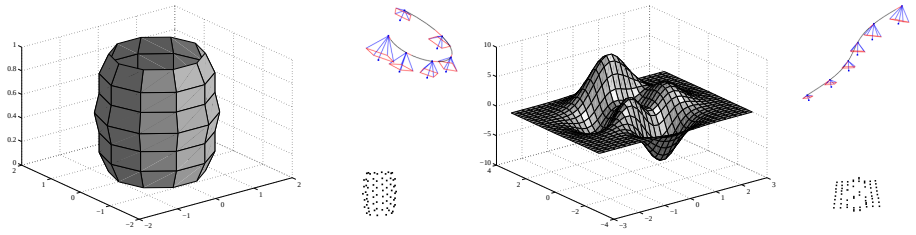


Fig. 3. Experiment 1. Synthetic objects. The shaded plots (a) illustrate the shapes of the 3D objects. Data used in the experiments are just the 3D vertex coordinates and synthetic camera positions, typical examples of which are shown in the wireframe plots (b).

4.1 High-relief bodies, freely varying camera parameters

The first experiment is simply a test of the theory, in a case where static-scene reconstruction cannot compute Euclidean structure and motion. Two synthetic objects were created, a unit barrel and a "peaks" function, illustrated in Figure 3. A synthetic sequence of 25 images was generated. In each frame, the 3D points of each object had an independent Euclidean transformation applied, with typical camera motions as illustrated. Also at each frame, all camera parameters were randomly chosen, so that no parameter was constant through the sequence. The 3D points were projected to generate 2D tracks.

From the 2D tracks, a projective reconstruction $\{ {}^{\beta}\mathsf{P}_v \}_{v=1}^{\mathrm{V}}$ is computed—in an arbitrary projective frame—for each body. In a static scene, such a projective reconstruction is the best that can be achieved. In the multibody case, the DIAC constraint (7) provides

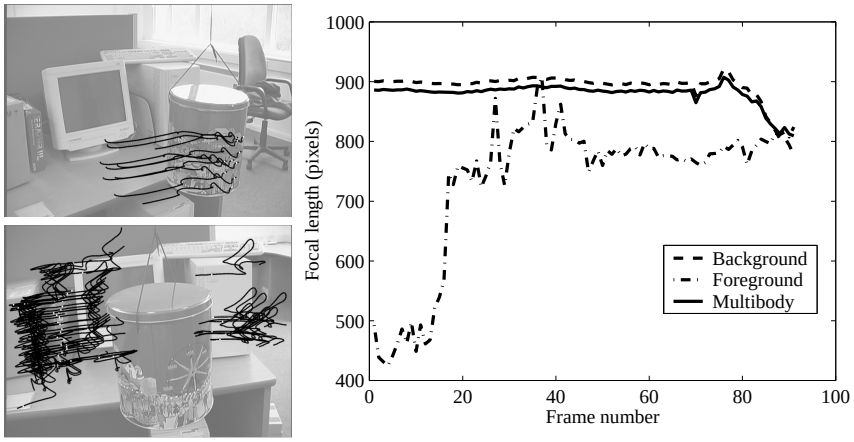


Fig. 4. Experiment 2. *Office sequence.* (a) Above: First frame of 90, with foreground object tracks superimposed. Below: Last frame, background object tracks superimposed. (b) Results of focal length estimation. Without multibody estimation, the foreground object has significant errors in focal length.

a way to compute a Euclidean upgrade. Given Euclideanizing homographies (3) for each structure, we can compare the DIACs of (7) using the distance measure of [18]:

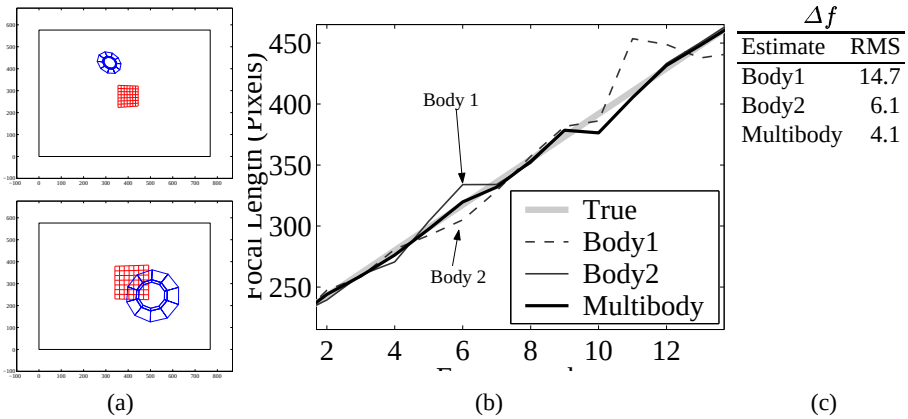
$$\epsilon = \sum_{v \in \text{VIEWS}} d^2\left({}^{\alpha}\boldsymbol{\omega}_v^*,\, {}^{\beta}\boldsymbol{\omega}_v^*\right) = \sum_{v \in \text{VIEWS}} \left\| \frac{{}^{\alpha}\boldsymbol{\omega}_v^*}{\|\,{}^{\alpha}\boldsymbol{\omega}_v^*\|_F} - \frac{{}^{\beta}\boldsymbol{\omega}_v^*}{\|\,{}^{\beta}\boldsymbol{\omega}_v^*\|_F} \right\|_F^2$$

The Euclideanizing homographies are parametrized in terms of K_1, $\mathbf{v}_\alpha$, $\mathbf{v}_\beta$, and a non-linear minimization of the above distance measure was implemented. Initializing the minimization from various sites around the solution converged, as expected, to the true solution. Note that in this experiment, a static-scene reconstruction converges to a different (and erroneous) solution as the problem is underconstrained.

4.2 Office sequence

Figure 4 shows a camcorder sequence of two objects: a static office background and a rotating cake tin. The camera was moved through the scene, without zooming, but with autofocus enabled. A projective reconstruction was computed for each object [9]. Then the reconstruction was used to initialize a bundle adjustment in which the focal length was allowed to vary. This provides a measure of reconstruction accuracy, as the recovered focal lengths should all be equal.

The graph shows the focal lengths computed through the sequence for (a) the background alone, (b) the tin alone, and (c) both simultaneously. In this case, the background structure has good relief, and the foreground object is quite shallow, so the combined estimate is very similar to the background. The foreground object, on the other hand,



(a) (b)

Δf

Estimate	RMS
Body1	14.7
Body2	6.1
Multibody	4.1

(c)

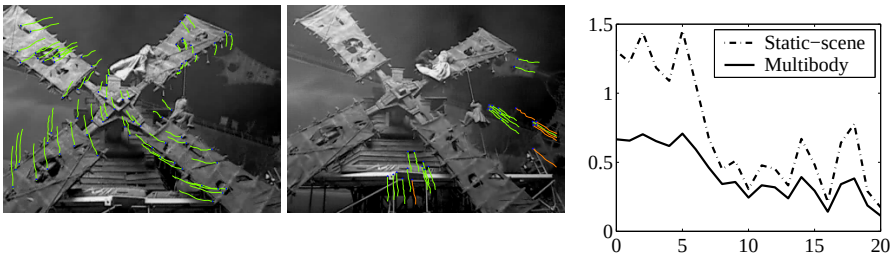
Fig. 5. Experiment 3. *Weak relief*. (a) Frames from the synthetic sequence. (b) Recovered focal lengths, plotted against frame number. (c) RMS difference between Δf and nominal value. The multibody estimate is closer to the ground truth and has lower variance than either of the single-object results.

enjoys a greatly improved calibration estimate. With only single-body estimation there are several equivalent reconstructions of the foreground points, corresponding to different choices of focal lengths. Without some other knowledge, such as that provided by multibody estimation, there is no way to choose the correct one.

4.3 Weak relief

In this synthetic experiment, we consider a case where a Euclidean reconstruction is possible for each moving object, but the solution is poorly conditioned. In this case, we expect that the combined estimation will improve both the precision and veridicality of the recovered reconstruction.

In this case, camera zoom was varied linearly, and the principal point was varied smoothly in a curve. Object motions were as in section 4.1 but the objects were scaled along one axis by a factor of 10 in order to reduce the relief. The 3D points were projected onto a synthetic image plane of size 768×576, and the 2D positions perturbed by noise vectors drawn from a 2D Gaussian distribution of mean zero and isotropic standard deviation $\sigma = 0.7$ pixels. A conventional single-scene reconstruction was computed for each object, finding the maximum likelihood estimate (9) via bundle adjustment (§3.2). A multibody estimate was computed by bundle adjustment, using the averaged camera parameters (§3.3) as a starting value. The camera focal lengths were extracted, and are shown in figure 5. As the focal length f is increasing linearly, the difference in f from frame to frame should be constant. The RMS difference between the algorithms' reported f and the nominal value was computed for each method, and the results are tabulated in figure 5. The multibody solution is, as expected, better than either of the static-scene estimates. (Simply averaging the focal lengths produces



Frame 0, Object A tracks Frame 20, Object B tracks Height vs. Frame number

Fig. 6. Experiment 4. *Planes*. Camera position fixed, object A is rotating, motion B arises due to zoom. Motion segmentation is automatic. The graph plots the Z coordinate of the computed optical centre (in the coordinate frame where the windmill is the plane $Z = 0$) against frame number, which should be constant through the reconstruction. Because the multibody estimate gains more accurate calibration parameters from the second motion, it has smaller standard deviation than the static-scene reconstruction using object A alone.

a worse RMS, obviously). The amount of improvement—RMS error reduces by a factor of 0.67—may appear small, but it is consistent with the use of twice as much data, which would imply an improvement of $1/\sqrt{2}$ or 0.71.

4.4 Planes

The fourth experiment uses a real sequence, some frames of which are shown in Figure 6. The camera is stationary, but zooming. The scene motion is automatically segmented (§3.1) into two components: the rotation of the windmill blades, and the homography induced by the zooming camera. Initial estimates of the camera parameters are obtained from the rotating plane. In this instance, not all points are tracked through all views, so there is missing data, but the bundle adjustment is easily modified to cope with this situation.

Figure 6 shows estimates of camera height (in windmill diameters) relative to the plane of the windmill before and after incorporation of multibody tracks. The height should be constant, so its standard deviation is a measure of the quality of the reconstruction. On this sequence, multibody analysis improves the standard deviation from 0.42 to 0.19, again demonstrating the advantage of considering multiple motions.

5 Discussion and Open Questions

We have extended the recovery of structure and motion to the case of sequences comprising multiple motions. We have presented an extension to static-scene bundle adjustment, which allows multiple motions to contribute to the estimation of camera parameters. We have shown that multibody reconstruction allows the analysis of scenes which cannot be handled by static-scene structure from motions, and that multibody analysis improves the accuracy of reconstructions even when standard techniques do apply. The

paper also contributes the new result that structure and motion can be determined from just four points in general position, even with one varying camera parameter – a result which can find applications in tracking through long sequences.

Although we have demonstrated in principle that certain problems can be solved, the solutions have been presented as nonlinear minimizations. It remains of interest to elucidate more elegant closed-form solutions to these problems, in particular:

1. Obtain the principal point and two focal lengths for the two view example of the introduction, based on the Krupa equations;
2. Determine a fixed calibration matrix and motion from four points imaged in sufficiently many views.
3. Obtain affine camera specializations for the scenarios and results of section 2.

References

1. M. Armstrong, A. Zisserman, and R. Hartley. Self-calibration from image triplets. In *Proc. ECCV*, LNCS 1064/5, pages 3–16. Springer-Verlag, 1996.
2. S. Avidan and A. Shashua. Trajectory triangulation of lines: Reconstruction of a 3D point moving along a line from a monocular image sequence. In *Proc. CVPR*, pages II:62–66, 1999.
3. M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proc. ICCV*, pages 231–236, 1993.
4. S. Bougnoux. From Projective to Euclidean space under any practical situation, a criticism of self-calibration. In *Proc. ICCV*, pages 790–796, Jan 1998.
5. J. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *IJCV*, 29(3):159–179, 1998.
6. C. Debrunner and N. Ahuja. Motion and structure factorization and segmentation of long multiple motion image sequences. In G. Sandini, editor, *Proc. ECCV*, pages 217–221. Springer-Verlag, 1992.
7. O. D. Faugeras. *Three-Dimensional Computer Vision: a Geometric Viewpoint.* MIT Press, 1993.
8. O. D. Faugeras, Q. Luong, and S. Maybank. Camera self-calibration: Theory and experiments. In *Proc. ECCV*, LNCS 588, pages 321–334. Springer-Verlag, 1992.
9. A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. ECCV*, pages 311–326. Springer-Verlag, Jun 1998.
10. C. W. Gear. Multibody grouping from motion images. *IJCV*, 29(2):133–150, 1998.
11. C. J. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conf.*, pages 147–151, 1988.
12. R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *Proc. ECCV*, LNCS 588, pages 579–587. Springer-Verlag, 1992.
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 and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, ISBN: 0521623049, 2000.
15. A. Heyden and K. Astrom. Flexible calibration: Minimal cases for auto-calibration. In *Proc. ICCV*, pages 350–355, 1999.

16. G. Newsam, D. Huynh, M. Brooks, and H. Pan. Recovering unknown focal lengths in self-calibration: an essentially linear algorithm and degenerate configurations. In *Proc. 18th ISPRS Congress, (International Archives of Photogrammetry and Remote Sensing (IAPRS), vol. XXXI)*, 1996.
17. M. Pollefeys. *Self-calibration and metric 3D reconstruction from uncalibrated image sequences.* PhD thesis, ESAT-PSI, K.U.Leuven, 1999.
18. 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.
19. M. Pollefeys, L. Van Gool, and A. Oosterlinck. The modulus constraint: a new constraint for self-calibration. In *Proc. ICPR*, pages 31–42, 1996.
20. A. Shashua, S. Avidan, and M. Werman. Trajectory triangulation over conic sections. In *Proc. ICCV*, pages 330–336, 1999.
21. D. Sinclair. Articulation detection for locally rigid objects. In R. Mohr and C. Wu, editors, *Europe-China Workshop on Geometrical Modelling and Invariants for Computer Vision*, pages 136–142. Xidan University Press, 1995.
22. P. Sturm. Critical motion sequences for the self-calibration of cameras and stereo systems with variable focal length. In *Proc. BMVC*, 1999.
23. P. H. S. Torr. *Motion segmentation and outlier detection.* PhD thesis, Dept. of Engineering Science, University of Oxford, 1995.
24. 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, Jan 1998.
25. 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.
26. P. H. S. Torr, A. Zisserman, and D. W. Murray. Motion clustering using the trilinear constraint over three views. In R. Mohr and C. Wu, editors, *Europe-China Workshop on Geometrical Modelling and Invariants for Computer Vision*, pages 118–125. Xidan University Press, 1995.
27. W. Triggs. Auto-calibration and the absolute quadric. In *Proc. CVPR*, pages 609–614, 1997.
28. W. Triggs. Autocalibration from planar scenes. In *Proc. ECCV*, 1998.
29. T. Viéville and D. Lingrand. Using singular displacements for uncalibrated monocular vision systems. In *Proc. ECCV*, LNCS 1064/1065, pages 207–216. Springer-Verlag, 1996.
30. R. G. Willson and S. A. Shafer. What is the center of the image? *J. Opt. Soc. Am.*, 11(11):2946–2955, Nov 1994.