Two-View Multibody Structure-and-Motion with Outliers through Model Selection

Konrad Schindler, *Member*, *IEEE*, and David Suter, *Senior Member*, *IEEE*

Abstract—Multibody structure-and-motion (MSaM) is the problem to establish the multiple-view geometry of several views of a 3D scene taken at different times, where the scene consists of multiple rigid objects moving relative to each other. We examine the case of two views. The setting is the following: Given are a set of corresponding image points in two images, which originate from an unknown number of moving scene objects, each giving rise to a motion model. Furthermore, the measurement noise is unknown, and there are a number of gross errors, which are outliers to all models. The task is to find an optimal set of motion models for the measurements. It is solved through Monte-Carlo sampling, careful statistical analysis of the sampled set of motion models, and simultaneous selection of multiple motion models to best explain the measurements. The framework is not restricted to any particular model selection mechanism because it is developed from a Bayesian viewpoint: Different model selection criteria are seen as different priors for the set of moving objects, which allow one to bias the selection procedure for different purposes.

Index Terms—Dynamic scenes, structure-and-motion, model selection, 3D motion segmentation.

1 Introduction

In the last decade, structure-and-motion recovery from perspective images as the only source of information has been extensively studied in the computer vision community. For the case of static scenes, the problem of fitting a 3D scene compatible with the images is well understood and essentially solved. There is a vast body of literature, from the pioneering works of Longuet-Higgins [21], Faugeras [6], and Hartley [11] to the comprehensive theory now presented in several excellent textbooks [5], [13], [22]. A key result is that, given a number of corresponding points, two images are enough to recover the 3D scene structure and the relative camera positions up to a projective transformation. Furthermore, it turned out that the type of geometric relation between corresponding points depends on the scene structure and on the relative camera motion. Not all scenes and not all relative camera positions can be appropriately described by the most general model, the epipolar geometry, encoded algebraically by the *fundamental matrix*. There are several cases in which the fundamental matrix becomes degenerate and must be replaced by a more restrictive model [5]. If either the camera motion is a pure rotation, or the scene is planar, then the relation between the two images is a projectivity, algebraically expressed as a *homography*. If the perspective distortion is small due to small motion or large focal length, it may be more appropriate to use an affine fundamental matrix or an affinity. To decide between different types of motions, a suitable model selection criterion is needed, which balances goodness-of-fit against model complexity. The first application of model selection to two-view motion models is due to Kanatani [18], who also first recognized that the dimension of the fitted manifold requires separate treatment [17].

Soon after the main SaM-theory had been established, researchers turned to the more challenging case of *dynamic* scenes, where the segmentation into independently moving objects and the motion estimation for each object have to be solved simultaneously. Even in the case of rigidly moving scene parts, which we will call *multibody structure-and-motion* or MSaM, the geometric properties of dynamic scenes turned out to be nontrivial [2], [10], [30], [32], [45]. Recently, an excellent extension of algebraic two-view SaM-theory to dynamic scenes has been presented [39], [40]. The theory is based on the assumption that each image measurement is explained by one out of a collection of fundamental matrices (termed the "multibody fundamental matrix"). The original method has been extended to the case of multiple homographies [38], and the same line of research has also been used to tackle the model selection problem [14]. The underlying mechanism, "minimum effective dimension," by definition aims to reduce the model dimensions as long as the goodness-of-fit does not drastically deteriorate. For example, it would prefer to explain a scene as six independently moving planes, rather than a single moving cube, which is somewhat counter-intuitive. Also, the purely algebraic approach does not allow for an outlier model, which, together with the nonlinear nature of the problem, makes it potentially vulnerable to gross measurement errors.

A different way to tackle the two-view MSaM problem is not to extend the geometric model, but instead try to cluster the points according to their motions. This leads to a chicken-and-egg problem: The motion models are needed for clustering, but the clustering is needed to compute the motion models. Irani and Anandan have proposed an iterative method to recover a number of homographies describing the scene [16]: The most dominant homography is extracted and the points consistent with it are removed, until the whole scene is explained. An iterative scheme, which is somewhere between simultaneous and iterative methods, has been proposed by Tong et al. [34] to extract multiple

• *The authors are with the Electrical and Computer Systems Engineering Department, Monash University, Clayton Campus, Wellington Road, 3800 VIC, Australia.*
E-mail: {konrad.schindler, d.suter}@eng.monash.edu.au.

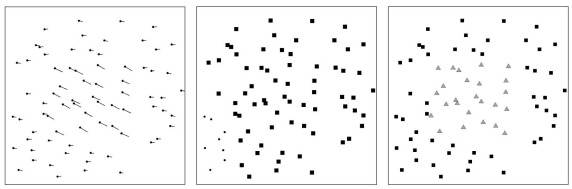
Manuscript received 17 Jan. 2005; revised 21 Apr. 2005; accepted 19 Oct. 2005; published online 13 Apr. 2006.
Recommended for acceptance by P.H.S. Torr.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0035-0105.

fundamental matrices: Tensor-voting is used to separate the outliers from those points, which are inliers to *any* epipolar geometry, then the single fundamental matrices are extracted with iterative RANSAC. Torr has proposed an iterative strategy for a combination of different motion models [35]: A single motion is estimated, the points consistent with it are removed from the data, then the next motion is estimated. In this scheme, each cluster is detected independently, disregarding the presence of other clusters in the data. The motion models are disjoint and their likelihood can be directly summed, producing a new model selection criterion.

The method presented here follows a recover-and-select scheme. In a first step, motion models are instantiated by Monte-Carlo sampling from the observed correspondences. Robust, nonparametric statistical analysis of the residuals is used to individually estimate the scale of the noise for each model. Note that the method is based on the (scalar) fitting residuals and, thus, not tied to any particular type of model. It is possible to use only epipolar geometries, only projectivities, only affinities, or any combination. In the following, we will assume that the effects of perspective projection are noticeable and only consider full fundamental matrices and homographies, but the framework is general and can be extended to other motion models, as, for example, shown in [35]. After the scale of the noise and the number of inliers for this scale have been found for every putative motion model, the likelihood of the motion can be computed. Given the likelihood as a measure for the goodness-of-fit, an optimal set of motion models can be selected from the candidate set with geometric model selection. Again, the presented framework is not tied to any particular model selection criterion and, in fact, we will argue that it is not possible to devise an all-purpose criterion for the whole range of possible applications.

There are two original contributions in this paper, one in each step. First, other than previous applications of geometric model selection, the presented method estimates the scale of the noise from the data. Compared with a globally preset threshold, this improves the capability to discriminate between different tentative motion models: A global threshold for inlier/outlier separation does not take into account the shape of the actual residual distributions, and therewith obscures the statistical properties of the data: If the threshold is higher than the width of the distribution, then the number of inliers and the standard deviation are overestimated; if the threshold is too low, the two quantities are underestimated. The incorrect estimates will influence model selection because these quantities are exactly the variables used to assess the goodness-of-fit. In contrast, the present method recovers the residual distribution for each tentative motion and estimates an individual standard deviation from it.

Second, previous iterative approaches to outlier-tolerant MSaM tacitly regard the candidate motion models as statistically independent, which is clearly not true since they may overlap (i.e., there are points which satisfy more than one motion). Iterative MSaM will assign such points to the motion detected first, rather than to the one they are most likely to belong to. This not only influences the classification of certain points (which can be remedied through postprocessing), but also the selection of the motions themselves because the inclusion or exclusion of a certain motion influences the likelihood of others. An example where iterative MSaM fails is shown in Fig. 1. This paper demonstrates simultaneous selection of all motion models. A new formulation for the posterior likelihood is derived, which properly accounts for the joint likelihood between overlapping motions. Selecting a set of motions and finding their respective inliers becomes a one-shot procedure.



(a) (b) (c)

Fig. 1. Iterative versus simultaneous motion recovery. (a) Disparities for a scene with two planar objects. The camera moves to the left, while the smaller object in the center moves down and to the right. (b) Result with iterative MSaM. The homography with the highest individual likelihood is a false motion, which explains most of the image points. (c) Result with simultaneous MSaM. The combination of two homographies with lower standard deviation has a higher likelihood, although their individual likelihoods are lower. Both examples use the same candidate set.

The paper is structured as follows: Section 2 describes how candidate motion models are blindly estimated from the data. In Section 3, the principle of geometric model selection is introduced and, on this grounds, an objective function for simultaneous selection of multiple motion models is derived. It is shown how this function can be optimized and how its behavior can be influenced with a prior which alters the penalty for model complexity. Section 4 puts these elements together to obtain a work-flow for MSaM. Experimental results with both synthetic and real data are presented in Section 5 and Section 6 gives a short summary and discussion.

1.1 Preliminaries and Notation

This section introduces the notation and recalls some basic elements of classical structure-and-motion theory, which are used as a basis for our multibody structure-and-motion algorithm. Points in the 2D image plane of a camera are represented by homogeneous 3-vectors $\mathbf{p} = w[x, y, 1]^\top$. Two projective cameras see the same 3D point as corresponding image points $\mathbf{p}_1$ and $\mathbf{p}_2$. For general camera motion and scene geometry, the two corresponding points are related by the epipolar geometry, which is algebraically expressed by a (3×3) fundamental matrix, such that

$$\mathbf{p}_2^\top \mathbf{F} \mathbf{p}_1 = 0. \tag{1}$$

The fundamental matrix has seven degrees of freedom and can be estimated from a set of seven correspondences using Hartley's seven-point algorithm [12]. For ≥ 8 correspondences, the eight-point algorithm offers a linear solution for $\mathbf{F}$, however the optimal estimator is nonlinear. The estimation procedure can be viewed as a geometric fitting problem: The coordinates of corresponding image points $\widetilde{\mathbf{p}} = [x_1, y_1, x_2, y_2]^\top$ are points in a 4D space, and the fundamental matrix is a 3D manifold, which needs to be fitted to a number of such points. Since the image point measurements are corrupted by noise, a correspondence $\widetilde{\mathbf{p}}$ will not lie exactly on the fundamental matrix, but will differ from it by a residual ϵ. To quantify the residual, we use the first-order approximation of the geometric error in the image plane, the so-called "Sampson distance" (but any

other geometrically meaningful error-measure could be used). The Sampson distance ϵ is given by [13]

$$\epsilon^2 = \frac{\left(\mathbf{p}_2^\top \mathtt{F}\mathbf{p}_1\right)^2}{\left(\mathtt{F}\mathbf{p}_1\right)_1^2 + \left(\mathtt{F}\mathbf{p}_1\right)_2^2 + \left(\mathtt{F}^\top\mathbf{p}_2\right)_1^2 + \left(\mathtt{F}^\top\mathbf{p}_2\right)_2^2}, \tag{2}$$

where $(\mathtt{F}\mathbf{p}_1)_1$ means the first element of the vector $\mathtt{F}\mathbf{p}_1$.

If the camera motion between the two views is a pure rotation around the projection center, or if the 3D points are all incident to a single plane R, then the correspondences are constrained to a projectivity: Point $\mathbf{p}_1$ can be directly transferred to point $\mathbf{p}_2$ by intersecting the corresponding ray with the second image plane, respectively by projecting it onto the scene plane and back-projecting the resulting scene point. The algebraic relation is a (3×3) matrix H called a homography, where

$$\mathbf{p}_2 \sim \mathtt{H}\mathbf{p}_1. \tag{3}$$

It has eight degrees of freedom, which can be determined from ≥ 4 correspondences by reordering the linear relation $\mathtt{H}\mathbf{p}_1 \times \mathbf{p}_2 = 0$ into an equation system $\mathtt{A}\mathbf{h} = 0$, such that $\mathbf{h}$ contains the nine unknown entries of H, and solving for $\mathbf{h}$. Again, the optimal estimator in nonlinear. In terms of geometric fitting, the homography is a 2D manifold in the 4D correspondence space, and the Sampson-distance with respect to a homography is given by

$$\epsilon^2 = \mathbf{h}^\top \mathtt{A}^\top (\mathtt{J}\mathtt{J}^\top)^{-1} \mathtt{A}\mathbf{h}, \tag{4}$$

where $\mathtt{J} = \frac{\partial(\mathtt{A}\mathbf{h})}{\partial(\tilde{\mathbf{p}})}$ is the Jacobian of the linear equation system.

From now on, the *type* of motion shall refer to the algebraic class of motion model without specifying parameters, e.g., the epipolar geometry is a type of motion and the projectivity is a different type of motion. A particular instance of a certain type, defined by its parameters, will be called a *motion* or *motion model*, so two fundamental matrices corresponding to different camera setups are different motions of the same type. The term *model* shall be reserved for the complete description of the data consisting of several motions of possibly different types. Recovering the multibody structure-and-motion is thus the task to explain the image measurements by fitting a model, which is a collection of motion models of variable type, where the set of available types is known, whereas the number of motions is unknown.

2 Generating Candidate Models

2.1 Sampling

For model selection, a set of candidate motions has to be generated. This is done with a simple Monte-Carlo procedure: Motion models are randomly instantiated from a minimal set of correspondences (seven for a fundamental matrix, four for a homography). Unfortunately, in a scene with multiple motions, only a comparatively small fraction of all correspondences belongs to each motion. Applying brute-force random sampling is already expensive if two motions are present and becomes intractable for more than two motions; for example, if we assume that the smallest inlier set comprises 20 percent of the data (an optimistic guess for three motions plus some outliers), the standard formula for RANSAC shows that we would need $\frac{\log(0.99)}{\log(1-0.2^7)} = 359,777$ samples to obtain an outlier-free sample with 99 percent confidence. Even if completely awkward samples are discarded at an early stage, this figure is an order of magnitude too high for practical applications.

Fig. 2. Local sampling scheme for tentative motion models. Samples are drawn from subregions of the image plane to exploit spatial coherence and reduce the required sample number.

A solution is to exploit the spatial coherence of points belonging to the same motion. Except for special cases such as transparent objects, points belonging to the same rigid object will be clustered in the image plane, and a local sampling scheme will therefore dramatically reduce the number of samples required to find an uncontaminated set. For the experiments in Section 5, the image plane was subdivided into three overlapping rows and three overlapping columns, and samples were drawn from the entire image, each column, each row, and each of the nine regions defined by a row-column intersection (see Fig. 2). This heuristic subdivision scheme proved to be a reasonable compromise between local coherence and global extent, which works well for different images. To justify the plausibility of the heuristics, we may say the following: On one hand, one column-row intersection in the scheme covers 11 percent of the image plane. Hence, if an independently moving object covers at least 10 percent of the image and is not very elongated in shape, there will be at least one region in which the object occupies ≈ 50 percent of the entire area, which (except for outliers) requires < 600 samples per region. On the other hand, the larger regions help to obtain a better distribution of the sampled points on large objects.

2.2 Estimating Standard Deviations

Given a motion model and a number of data points, the scale of the noise can be estimated without any further knowledge by applying the TSSE-estimator of Wang and Suter [43]. To this end, the residuals of all data points with respect to the motion model have to be computed (we use Sampson-residuals, see Section 1.1). Mean-shift analysis [4] of the ordered absolute residuals yields a nonparametric estimate for their probability density function. Assuming that the inliers have mean zero, the valley of this function, which is closest to 0, is a sensible point to separate inliers from outliers. The bandwidth for the mean-shift algorithm can be selected automatically from the data with an oversmoothed bandwidth selector [42], [44], so that the procedure does not involve any manually chosen parameters. The procedure is illustrated in Fig. 3. Further examples are shown in Figs. 6d, 6e, and 6f. Note that the underlying assumptions are very weak: Rather than assuming any particular distribution, we only assume that the residuals of the inliers should have zero-mean and that their distribution is symmetric (since we use the absolute residuals).

As widely known in the statistical literature, e.g., [26] and also noticed by computer vision researchers [46], the *efficiency* of random sampling methods is poor, i.e., even a motion constructed from the best uncontaminated random sample may differ quite strongly from the optimal fit. Therefore, it is necessary to refine each tentative motion with a least-squares fit to the inlier points.

Estimating the inlier threshold and variance of each motion separately from the data considerably improves the power of

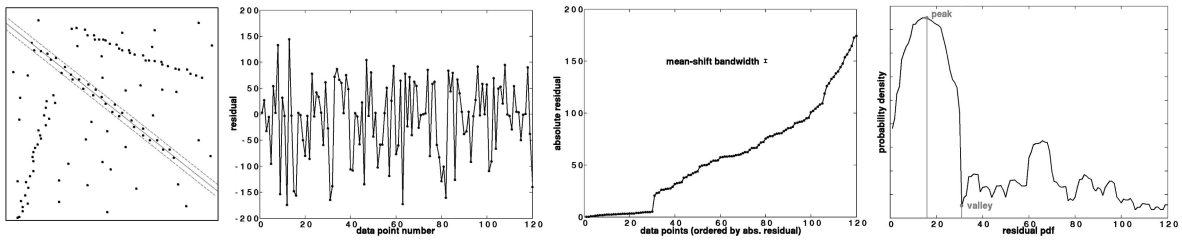


Fig. 3. Simultaneous scale estimation and outlier detection with the TSSE-estimator. (a) A set of data points containing several structures and some outliers, a (manually created) candidate for a straight line fit, and the estimated inlier/outlier boundary. (b) Residuals of all data points with respect to the line. (c) Ordered absolute residuals and kernel window with automatically selected bandwidth. (d) Detected peak and valley of the distribution.

the method, compared with a fixed threshold between inliers and outliers. When searching for a *single* motion, a slightly incorrect threshold is not problematic, while it may impair the results in the presence of multiple motions. There are two possible cases: If the threshold is too low, not all inliers are found; however, the motion is still fitted entirely to inliers, which will give a good result (in fact, some authors recommend this strategy to assure that no outliers compromise the fit, e.g., [7]). However, when searching for multiple motions, the situation is different. If only a subset of the inliers is found and assigned to the motion, the remaining inliers will give rise to a second model, leading to overfitting. If, on the contrary, the threshold is too high, it will still remove a large part of the outliers, so that, in the presence of a single motion, a robust least-squares technique such as an M-estimator [15] can be used to obtain a correct fit. Again, the situation is more complicated in the presence of multiple motions: If one motion (either the one with larger support or simply the one detected first) claims too many data points, it may weaken the support for a second model, to which the points actually belong. This can lead to underfitting. The two cases are schematically illustrated in Fig. 4.

3 Model Selection

3.1 Principle of Geometric Model Selection

To select the optimal set of motions, a criterion is needed, which balances the goodness-of-fit against the complexity of the complete description by penalizing the addition of new motion models. Both adding more motions and using a more complex motion type obviously decreases the total fitting error because degrees of freedom are added which allow the model to adapt better to the data. The basic idea of model selection is to counteract this behavior by assigning a cost to each model type, which grows with the dimension of the associated manifold and with the number of parameters required to define it.

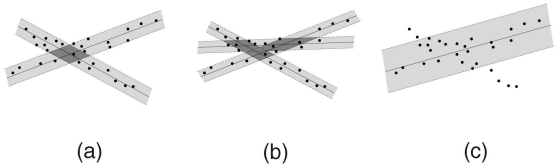


Fig. 4. Influence of wrong thresholds on fitting multiple straight lines. (a) Correct threshold and fitted lines. (b) Thresholds that are too low encourage overfitting: Data points missed by the fit give rise to another line with low residuals. (c) Thresholds that are too large encourage underfitting: Data points wrongly assigned to the fit weaken the support for other lines.

There are several criteria in the statistical literature, starting with Wallace's *minimum message length* MML [41]. The first simple and widely used criterion is Akaike's *an information criterion* AIC [1]. It is based on the Kullback-Leibler divergence and sets the complexity penalty such that the *future residual* is minimized. However, it has been criticized both theoretically (for not being asymptotically consistent) and empirically (for overfitting) because it does not account for the number of data points. Methods, which address this problem, are Schwartz' *Bayes information criterion* BIC [28], an approximation to maximizing likelihood that the data has been generated by the model, and Rissanen's *minimum description length* MDL [25], which is an information theoretic criterion similar to MML, and seeks to minimize the coding length of the data. Note that each criterion follows a different, but very restrictive, definition of optimality (best model for unobserved data, most probable model, most compact model).

In practice, all criteria have to use approximations, and in their standard form assume that the dimension of the fitted manifold is known and only the number of parameters of that manifold varies. Since we have to decide between motions of different dimension, an extension is needed—otherwise, the one with higher dimension will always be selected because it is less restrictive (e.g., the errors of any point cloud with respect to a straight line are lower or equal to the errors with respect to a point). In computer vision, this problem was first recognized by Kanatani, who solved it through an extension of AIC, called the *geometric information criterion* GIC[1] [17]. GIC selects the model $\mathcal{M}$ which maximizes

$$\mathrm{GIC}(\mathcal{M}) = 2\ln(\mathcal{L}) - 2(N_t D + K), \tag{5}$$

where N_t is the total number of correspondences, K is the number of parameters of the fitted manifold (eight for a homography, seven for a fundamental matrix), and D is the dimension of the manifold (two for a homography, three for a fundamental matrix). $\mathcal{L}$ is the likelihood of the model.

1. The literature is not consistent. On other occasions, the same criterion is referred to as G-AIC.

Matsunaga and Kanatani have also extended MDL to a geometric model selection criterion termed G-MDL [24], with the slightly different cost function

$$\text{G-MDL}(\mathcal{M}) = 2\ln(\mathcal{L}) - (N_t D + K)\ln(\sigma^2), \tag{6}$$

where σ is the noise level. A currently unresolved issue is that the criterion in its present form is not invariant to scaling with a scalar due to the dependence on σ. A heuristics has been proposed to remedy the scale-dependence [19]; however, we believe that this is still preliminary, and further research is needed to clarify the issue.

A similar extension for BIC, based on Bayesian decision theory, is the core of Torr's work on selecting motion models. His criterion is termed *geometrically robust information criterion* GRIC[2] [35], [37]. GRIC selects the model $\mathcal{M}$ which maximizes

$$\text{GRIC}(\mathcal{M}) = 2\ln(\mathcal{L}) - N_t D\ln(R) - K\ln(RN_t), \tag{7}$$

where R is the dimension of the data (4 for pairs of image points).

Several authors quite correctly make the point that there is no "canonical" way to select a model—choosing a model is an interpretation of the data, and the choice depends on the model's purpose [7], [18]. We agree with this view and, in fact, will show that one can construct a prior which converts one criterion into the other. In probabilistic terms, different complexity penalties correspond to different priors, which encode different expectations about the selected model. For specific tasks, these expectations may be quite different from the ones expressed by one of the standard model selection criteria. We feel that the Bayesian view most naturally fits into our probabilistic framework and will use GRIC in the rest of the paper; however, both the likelihood and the formulation of the optimization problem given in the following are generic and can just as well be used with GIC or G-MDL, changing only the penalty terms.

3.2 Computing the Likelihood

In order to compute the likelihood of a model, we first have to choose suitable probability distributions for the data points. We try to avoid any unjustified assumptions about the data and choose the least informative distributions. To find the least informative distribution consistent with some constraints, Shannon has introduced the principle of maximum entropy [29]. So far, we have assumed that the residuals of the inliers to a motion have zero-mean and are symmetrically distributed. One more assumption that we need is that the distribution is simple enough to discard the higher-order moments and characterize it by its second moment. In [3], it is shown that, if we base our estimates only on the first and second moments of the noise, the least informative distribution under Shannon's definition is a Gaussian. Let $\mathcal{V}_i$ denote a tentative motion model with standard deviation σ_i, and let $\mathbf{p}$ be a correspondence, which has the residual ϵ_i with respect to $\mathcal{V}_i$. Then, the likelihood of $\mathbf{p}$ is

$$\mathcal{L}(\mathbf{p}|\mathcal{V}_i) = \frac{1}{(\sigma_i\sqrt{2\pi})^4}\exp\left(-\frac{\epsilon_{(i),k}^2}{2\sigma_i^2}\right). \tag{8}$$

If we denote the set of all N_i inliers to $\mathcal{V}_i$ by $\{\mathbf{p}_k, k \in \mathcal{V}_i\}$ and their residuals with respect to $\mathcal{V}_i$ by $\epsilon_{(i),k}$, then the total likelihood of $\mathcal{V}_i$ is

$$\mathcal{L}_i = \prod_{k\in\mathcal{V}_i}\left(\frac{1}{(\sigma_i\sqrt{2\pi})^4}\exp\left(-\frac{\epsilon_{(i),k}^2}{2\sigma_i^2}\right)\right) = \prod_{k\in\mathcal{V}_i}\mathtt{G}_k^{(i)}. \tag{9}$$

Although the distinction may seem academic, it should be noted that this does *not* say that the noise actually is Gaussian, but only that it can be described well enough by its first and second moments.

In the same way, we only make the weakest possible assumptions about the outliers, which do not conform to any of the motion models. As expected, the least informative distribution for a value, for which we only know a lower and upper bound, is a uniform distribution. Hence, if the image plane of the first image has the area A_L (measured in the same unit as the residuals), and the search window within which a correspondence is searched in the second image has the area A_R, then the likelihood of point $\mathbf{p}$ being an outlier is

$$\mathcal{L}(\mathbf{p}|A_L, A_R) = \frac{1}{A_L A_R} = P. \tag{10}$$

Again, this does not say that the outliers are uniformly distributed, but only that all we know is the region of the plane in which they can possibly lie according to the employed matching procedure. If no spatial constraints are enforced during matching, then A_R is the entire area of the second image.

Since we want to select a subset of all motions established previously, the total likelihood has to be split into the contributions from the single motions. At the same time, we have to account for the fact that data points may be inliers to more than one candidate motion: If a data point has a sufficiently low residual in more than one model, it is not possible to determine reliably from the data which of the corresponding distributions it has been sampled from. We will call the set of points, which have low residuals with respect to two different motions, the *overlap* between the two motions.[3] We will from now on assume only pairwise overlap. This assumption is not strictly correct and causes overly large overlap penalties, if a point satisfies more than two motions, but the number of these points is small compared to those satisfying exactly two motions. The approximation is necessary to yield a tractable optimization problem, as explained later in this section.

Points in the overlap should contribute only once to the overall likelihood since a 3D point cannot lie on more than one of several physically disjoint objects. It is important to understand that correct treatment of motion overlap is a fundamental requirement in a scheme, which uses model selection to simultaneously recover multiple motions (as described in Section 3.3). If it is neglected, any motion whose likelihood outweighs the complexity penalty will increase the total likelihood and, thus, will be selected. As an extreme example, imagine the case that two identical motions were present in the candidate set (overlap 100 percent). Regarded on their own, both will have the same likelihood and, if this

2. Note: The same author has called the criterion GBIC in later work [36].

3. Note that this definition of overlap is only based on residuals and does not look at a point's position in the image plane.

likelihood is positive, they will both be selected, which clearly contradicts the desire to minimize the complexity of the data description—there is no benefit in "explaining the same point twice."

Let us first look at two tentative motions, $\mathcal{V}_i$ and $\mathcal{V}_j$. If both are used and they overlap, then a point in the overlap should only contribute to the motion it fits better.[4] Let $\{\mathbf{p}_k, k \in \mathcal{V}_{[ij]}\}$ denote the $N_{[ij]}$ points, which are inliers to both motions $\mathcal{V}_i$ and $\mathcal{V}_j$. Some part $\mathcal{V}_{[i]}$ of these points will have lower likelihood in $\mathcal{V}_i$, the remainder $\mathcal{V}_{[j]}$ will have lower likelihood in $\mathcal{V}_j$. If the two motions were regarded as independent, their joint likelihood would be $\mathcal{L}_{i\cup j} = \mathcal{L}_i\mathcal{L}_j$. In this expression, each point of the overlap also makes an unjustified contribution to the motion, where it has *lower* likelihood. If we call the total amount of these unjustified contributions $\mathcal{L}_{[ij]}$, the correct joint likelihood of the two motions is given by $\mathcal{L}_{i\cup j} = \frac{\mathcal{L}_i\mathcal{L}_j}{\mathcal{L}_{[ij]}}$, where

$$\mathcal{L}_{[ij]} = \prod_{k\in\mathcal{V}_{[ij]}} \min\left(\mathtt{G}_k^{(i)}, \mathtt{G}_k^{(j)}\right) = \prod_{k\in\mathcal{V}_{[i]}} \mathtt{G}_k^{(i)} \prod_{k\in\mathcal{V}_{[j]}} \mathtt{G}_k^{(j)}. \tag{11}$$

Let the set of all candidate motions (fundamental matrices and homographies) be $\mathcal{C} = \{\mathcal{V}_1 \ldots \mathcal{V}_M\}$. If we select a subset $\widehat{\mathcal{C}}$ of $\mathcal{C}$, then $\widehat{\mathcal{C}}$ will explain some of the correspondences $\mathbf{p}_k$ and leave the remaining H correspondences as outliers. According to (10), the total likelihood of the outliers is given by $\mathcal{L}_{/\widehat{\mathcal{C}}} = P^H$, and the total likelihood of the selected subset is

$$\mathcal{L}_{\widehat{\mathcal{C}}} = \mathcal{L}_{/\widehat{\mathcal{C}}} \frac{\prod_{i\in\widehat{\mathcal{C}}} \mathcal{L}_i}{\prod_{i,j\in\widehat{\mathcal{C}}} \mathcal{L}_{[ij]}}. \tag{12}$$

To compare different subsets $\widehat{\mathcal{C}}$, one can introduce a Boolean index vector $\mathbf{b}$ of length M (the number of candidate motions in $\mathcal{C}$), with elements $(b_i = 1)$ if motion $\mathcal{V}_i$ is used and $(b_i = 0)$ otherwise. Then, the log-likelihood is given by

$$\ln(\mathcal{L}) = \sum_{i\in\mathcal{C}} (b_i \ln(\mathcal{L}_i)) - \sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{C}} \left(b_i b_j \ln\left(\mathcal{L}_{[ij]}\right)\right) + H\ln(P). \tag{13}$$

In this expression, we can substitute the likelihoods with (9) and (11). Furthermore, we can express the number of outliers as the difference between the total number of points N_t and the number of inliers (again, assuming only pairwise overlap):

$$H\ln(P) = \left(N_t - \sum_{i\in\mathcal{C}} b_i N_i + \sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{C}} b_i b_j N_{[ij]}\right) \ln(P). \tag{14}$$

The constant term $N_t \ln(P)$ will not influence the optimization and can be dropped (by a slight abuse of notation, the new quantity is still called $\mathcal{L}$). Furthermore, we abbreviate the normalized sum of squared errors of a motion

$$\frac{1}{\sigma_i^2} \sum_{k\in\mathcal{V}_i} \epsilon_{(i),k}^2 = E_i \tag{15}$$

and the normalized sums of squared errors in the overlap between two motions

$$\frac{1}{\sigma_i^2} \sum_{k\in\mathcal{V}_{[i]}} \epsilon_{(i),k}^2 = E_{[i]}, \quad \frac{1}{\sigma_j^2} \sum_{k\in\mathcal{V}_{[j]}} \epsilon_{(j),k}^2 = E_{[j]} \tag{16}$$

so that

$$\begin{aligned} \ln(\mathcal{L}_i) &= 2N_i \ln(2\pi\sigma_i^2) + \frac{1}{2} E_i \\ \ln(\mathcal{L}_{[ij]}) &= 2(N_{[i]} \ln(2\pi\sigma_i^2) + N_{[j]} \ln(2\pi\sigma_j^2)) + \frac{1}{2}(E_{[i]} + E_{[j]}). \end{aligned} \tag{17}$$

Substituting these expressions in (13), multiplying by 2, setting $\lambda_1 = -2\ln(P) - 4\ln(2\pi)$, and reordering yields

$$\begin{aligned} 2\ln(\mathcal{L}) = &\sum_{i\in\mathcal{C}} (b_i(N_i\lambda_1 - 4N_i \ln(\sigma_i^2) - E_i)) - \\ &- \sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{C}} \Big(b_i b_j \big(N_{[ij]}\lambda_1 - 4N_{[i]} \ln(\sigma_i^2) \\ &- E_{[i]} - 4N_{[j]} \ln(\sigma_j^2) - E_{[j]}\big)\Big). \end{aligned} \tag{18}$$

In this form, the log-likelihood is only a function of the index vector $\mathbf{b}$. All other quantities on the right side are known parameters of the candidate motion models.

3.3 Maximizing the Criterion

Previously, model selection criteria have either been used to select one manifold of varying dimension at a time, such as in [17], [35], or to fit an unknown number of manifolds with the same dimension at once, such as in [20]. In that work, Leonardis et al. showed that one can formulate a tractable optimization problem for an unknown number of motions, if the contributions of different motions to the total likelihood can be separated. This is the reason why we assume only pairwise overlap.

With expression (18) for the likelihood, the GRIC (7) for a collection of motions $\widehat{\mathcal{C}}(\mathbf{b})$ can be written as a quadratic expression of the index vector:

$$\mathrm{GRIC}(\mathbf{b}) = \mathbf{b}^\top \mathtt{Q} \mathbf{b}, \tag{19}$$

where $\mathtt{Q}$ is a symmetric $(M \times M)$ matrix [20]. Let the constants $\lambda_2 = N_t \ln(4)$ and $\lambda_3 = \ln(4N_t)$. Then, the diagonal elements of $\mathtt{Q}$ are

$$q_{ii} = N_i\lambda_1 - 4N_i \ln(\sigma_i^2) - E_i - \lambda_2 D_i - \lambda_3 K_i. \tag{20}$$

The off-diagonal elements, which handle the overlap between different tentative motions, are

$$\begin{aligned} q_{ij} = q_{ji} = \\ -\frac{1}{2}\left(N_{[ij]}\lambda_1 - 4N_{[i]} \ln(\sigma_i^2) - E_{[i]} - 4N_{[j]} \ln(\sigma_j^2) - E_{[j]}\right). \end{aligned} \tag{21}$$

Intuitively, the cost function (19)

- favors motions which significantly reduce the number of outliers (large N_i),
- favors motions with small standard deviation and small fitting errors (low σ_i and low E_i),

4. Strictly speaking, this is not correct: Each point should contribute to the distribution it actually has been sampled from. Since there is no way to determine this in the case of overlapping distributions, we have to use the best guess.

- tries to keep the number of motions low (by penalizing each used motion $\propto K_i$), which is the usual complexity penalty for model selection.
- tries to keep the dimension of the motion low (by penalizing the motion $\propto D_i$), which is the extension to geometric model selection.

Note that no parameters have to be tuned in (20) and (21).

Maximizing expression (19) over $\mathbf{b}$ is a combinatorial problem for which a global optimum can only be found through exhaustive search. In operations research, Taboo-search [9] is a standard method for approximate solution of such problems. In spite of its simplicity and an obvious similarity to human searching behavior, it has rarely been used in computer vision, with the notable exception of [31]. In this work, Stricker and Leonardis advocate the use of Taboo-search for exactly the type of problem we have to solve and also report that it outperforms simulated annealing methods.

A detailed description of the method is beyond the scope of this paper, but we will briefly sketch the principle since it does not seem to be widely known in the computer vision community. Let us begin with a greedy search procedure: We start from an arbitrary set of motions (e.g., the empty set). The two possible elementary moves in a binary problem are to either switch on an additional motion or to switch off one of the currently used motions. The greedy solution iteratively searches the move with the highest benefit until no further reduction of the cost is possible. Taboo-search extends this method by *not* stopping at the first local minimum. Instead, the search continues, but recent moves are remembered in a "short-term memory," which is constantly updated. Undoing any of the moves in the memory is illegal ("taboo") unless it reduces the cost beyond the best current minimum. This ensures that the search departs far enough from a local minimum. Furthermore, the frequency of each move is remembered and, if all feasible moves are illegal, the least frequent one is chosen in order to diversify the search as much as possible without getting stuck. The best of the detected local minima is retained and, if it cannot be improved further in a preset number of iterations, the search is terminated.

As with most metaheuristics for hard optimization problems, the computational complexity and solution quality of Taboo-search depend on the implementation, and no provable guarantees can be given, except that the optimum will always be at least as good as the greedy result for the same search neighborhood. We have tested our implementation on a large number of random quadratic Boolean problems of varying size. Empirically, its average complexity is $\mathcal{O}(M^{2.5})$, with variations between different runs generally below 10 percent.

3.4 Constraints

For any real problem an upper bound for the allowable residual ϵ_{max} for a single point measurement is known—it is the distance above which a measurement is considered an "outlier" rather than a "noisy inlier." In the presence of a single motion, the maximum allowable residual would be a natural upper bound for the standard deviation σ of the motion model, since $\frac{1}{N}\sum \epsilon_k^2 \leq \max(\epsilon_k^2)$. A motion model with higher standard deviation is meaningless because it is at least partly based on points, whose residuals are too large to be inliers. As an extreme example, a motion with a standard deviation greater than half the image size will always explain *all* measurements within $\pm\sigma$. A constraint is needed to make sure that such meaningless motion models cannot be selected.

To account for outliers and pseudo-outliers on other motion models, which tend to blur the distinction between inliers and outlier, it is advisable to use a more conservative upper bound $t\epsilon_{max}, t \approx 2$. In order to formally add this constraint to the probabilistic formulation of the optimization problem, one would have to redefine the likelihood (9) of a candidate motion $\mathcal{V}_i$ as

$$\mathcal{L}_i = \begin{cases} \prod_{k\in\mathcal{V}_i} \left(\frac{1}{(\sigma_i\sqrt{2\pi})^4} \exp\left(-\frac{\epsilon_{(i),k}^2}{2\sigma_i^2} \right) \right) & \text{if } \sigma_i \leq t\epsilon_{max} \\ 0 & \text{else,} \end{cases} \tag{22}$$

which will give motions with too high σ_i an infinitely high goodness-of-fit penalty. Since the constraint is independent of the other terms of the objective function, using (22) is equivalent to removing motion models with $\sigma_i > t\epsilon_{max}$ from the candidate set prior to selection. The latter speeds up the optimization.

Note the difference to methods which threshold the residuals at the fitting stage (e.g., basic RANSAC): These methods determine the inlier set by discarding points with large residuals and, in this way, always obtain a fit with sufficiently low σ, but the boundary is arbitrary and may not be apparent in the corresponding probability density function. On the contrary, TSSE bases the partitioning into inliers and outliers on probability densities and, by definition, finds a boundary which is observable in the *pdf*. The threshold on σ is then used to discard the entire fit, rather than only some of its points.

3.5 Model Selection and Priors

As already stated earlier, choosing a model is an *interpretation* of the data, and the best solution may vary depending on the task at hand. Specifically, none of the given criteria gives satisfactory results if the task is to segment small relative motions. The issue is related to the definition of what is a "satisfactory" result: The purpose is not merely a compact description with low errors, but the discrimination of motions, which can be explained well enough with a single motion model. So, in some sense, we are *aiming for an overfit*. To bias model selection in the desired way, we only have to decrease the cost for a motion model, and the selection mechanism will automatically choose more motions with lower residuals and, in this way, separate similar motions. In fact, the reason why established model-scoring methods often fail in practice is that they are based on general definitions of optimality, which may not be suitable for the application. On the contrary, a Bayesian view of the problem allows one to impose a problem-specific definition of optimality: A prior on the set of motion models can be used to smoothly move the emphasis from greater sensitivity to greater compactness of the description, including the penalties of GIC, G-MDL, and GRIC as special cases.

In a Bayesian framework, information not manifest in the data is introduced in the form of the prior distribution. The penalty terms in the criterion express the belief that a more complicated description of the data is less likely. For the example of separating small motions, the prior shall mitigate this, saying that it is "not that much less likely." The prior must also be proportional to the total number of matches N_t;

otherwise, its influence will decrease $\rightarrow 0$ as the number of matches increases.[5] A simple prior with these properties is

$$\mathcal{L}_{\mathcal{P}} = \frac{1}{S_{\mathcal{P}}} \prod_{i \in \widehat{\mathcal{C}}} U^{B_i N_t}, B_i = \begin{cases} \mathtt{H} : 1 \\ \mathtt{F} : \frac{3\ln(4)N_t + 7\ln(4N_t)}{2\ln(4)N_t + 8\ln(4N_t)} \end{cases}. \quad (23)$$

$S_{\mathcal{P}}$ is the combinatorial sum over all possible $\prod U^{B_i N_t}$, which normalizes the total probability to 1, but it need not be known because it is constant and can be dropped. The factor B_i is introduced to preserve the ratio between the total penalties for a fundamental matrix and a homography, so that the prior does not bias the selection of the model type. The constant U determines the strength of the bias. Being part of the prior, it cannot be determined within the framework, but is an as yet arbitrary parameter, the choice of which requires external knowledge. Given that the model cost should be decreased, but remain > 0, the theoretical range is $(1 < U < 4^2)$. Writing $\lambda_4 = N_t \ln(U)$, the prior changes the diagonal elements of $\mathtt{Q}$ to

$$q_{ii} = N_i\lambda_1 - 4N_i \ln(\sigma_i^2) - E_i - \lambda_2 D_i - \lambda_3 K_i + B_i\lambda_4. \quad (24)$$

As desired, the penalties for adding motion models have been decreased, treating all motions in an equal way independent of the total number, and preserving the ratio between model type penalties. In Section 5, the effect of this prior is shown on a practical example.

The prior likelihood (23) is only the simplest representative of a more general prior

$$\mathcal{L}_{Pr} = \frac{1}{S_{Pr}} \prod_{i \in \widehat{\mathcal{C}}} U^{f(N_t)}, \quad (25)$$

where $f(N_t)$ is some function of N_t. The general form no longer treats all models equally, and it also allows one to influence the likelihood ratio between different motion types. For example, setting

$$f(N_t) = \begin{cases} \mathtt{H} : 2(\ln(4) - 2)N_t + 8\ln(4N_t) - 16 \\ \mathtt{F} : 3(\ln(4) - 2)N_t + 7\ln(4N_t) - 14 \end{cases} \quad (26)$$

results in a prior, which converts GRIC into GIC. In statistical theory, it has been pointed out that different model scoring methods can be seen as different priors [8]. We would like to emphasize that this is also true for geometric model selection, and that the established methods belong to a family defined by the function $f(N_t)$. In practice, it may be useful to go beyond the established penalty terms and design new scoring methods by changing the function f. However, it remains to be investigated how this could be done in a theoretically justified way. We do not recommend the use of arbitrary priors without clear interpretation, which are just the infamous "damping factors" in Bayesian disguise.

4 The Complete MSaM Algorithm

In the previous sections, the ingredients for a robust MSaM method have been developed. Putting them together yields a complete work-flow:

1. **Matching**. Obtain a set of corresponding points between the two images with a suitable algorithm. This step is not the topic of the present paper. For the experiments in this paper, we used manually measured correspondences, point tracks obtained with the publicly available implementation of the KLT-tracker [33], and wide-baseline matching with the *maximally stable extremal regions* (MSERs) of Matas et al. [23].
2. **Sampling**. Randomly sample a sufficient number of candidates for each considered model type, with the local scheme described in Section 2. If we assume an inlier fraction of at least 50 percent in one of the subregions, an easy calculation shows that we need to sample ≈ 600 fundamental matrices and ≈ 80 homographies per subregion to obtain an uncontaminated sample with a probability > 99 percent.
3. **Data analysis**. Estimate the standard deviation and the inlier set of each candidate with TSSE; as described in Section 2.2, refine the candidates with a least-squares fit and discard all candidates which do not satisfy the ϵ_{max}-constraint in (22).
4. **Optimization**. The result of the previous steps is a set of M candidate models (fundamental matrices and homographies). Each candidate consists of a known model type and estimates for the parameters, inlier count, standard deviation, and residuals. With these elements, build the matrix $\mathtt{Q}$ using (20) and maximize the objective function (19) with Taboo-search.
5. **Result**. The solution vector $\widehat{\mathbf{b}}$ directly gives the multibody structure-and-motion: The candidate models for which $(b_i = 1)$ are the ones which optimally explain the data, their respective inlier sets are the (nonexclusive) segmentation of the correspondences into different 3D motions, and the data points which are not in any of those inlier sets are the outliers.

5 Experiments

5.1 Simulations with Synthetic Data

Experiments with synthetic data were used to empirically assess the proposed method. The experiments assume a pair of images with 500×500 pixels. For the first experiment, spatially clustered clouds of 50 random points per model were generated on 1-3 randomly chosen motion models and perturbed with 0.5 pixel i.i.d. Gaussian noise. The amount of motion was chosen at random in such a way that all image points come to lie within the image boundaries, point clouds from different motions do not overlap, and extreme motions in depth are avoided (by limiting the expansion factor of a point cloud to $0.5 < E < 2$). Fifty outliers were added from a uniform distribution over the two image planes. The algorithm was applied to 100 such random data sets. To judge the performance of the selection, the number and the types of recovered motions are recorded; to judge the accuracy of the results, the number of inliers per motion and its standard deviation are used. The results of the experiment are given in Table 1. As expected, the estimates for the motions' standard deviations grow as more motions are added, since pseudo-outliers from other motion models blur the borders between the distributions. In some cases, one out of three motions was missed. This happens when two of the random motions are very similar and have a large overlap, so that the cost for assigning the remaining points of the

5. GRIC, as well as GIC and G-MDL, can only be evaluated for given N_t. Hence, the problem is to fit a set of motions to a *known* number N_t of a priori *unknown* correspondences, and N_t is indeed part of the prior knowledge.

TABLE 1
Three-Dimensional Segmentation of Random Data

number	detected	correct	inliers	σ of inliers [px]
1	100.0%	100.0%	49.8	0.56
2	100.0%	100.0%	50.3	0.69
3	90.6%	90.6%	51.8	0.77
1-3	*95.4%*	*95.4%*	*50.9*	*0.70*

Left to right: True number of motions, percentage of true motions detected in 100 runs, percentage of true motions which were assigned the correct model, average number of inliers (ground truth: 50), average standard deviation (ground truth: 0.5).

weaker one to the outliers is lower than the cost for an additional motion. This effect is inevitable in the presence of outliers: Allowing for unexplained points inherently reduces the ability to discriminate similar motions. The effect could be mitigated by a prior, which increases the cost of outliers—at the expense of spurious models in case of many outliers. All detected motions were assigned the correct motion model.

In a second set of experiments, the sensitivity to noise was assessed. For each test, two random motions were created with 50 inliers each and augmented with 50 outliers. The amount of noise added to the inliers was increased from 0.05 to 2.5 pixels (the minimal noise of 0.05 is required for correct bandwidth-selection during the mean-shift procedure). Thirty tests were run at each noise level. Since the ability to separate the two inlier distributions depends on the amount of outliers, the whole test was also repeated with 25 outliers. The results are shown in Fig. 5. Up to a noise level of 1.25 pixels (0.25 percent of the image size), the performance is stable, then it rapidly breaks down: The inlier distributions become increasingly wider and flatter and are no longer separable. The results with fewer outliers are slightly better, but support the conclusion that the method can handle up to $\approx 0.25\%$ noise.

The third experiment again used two random motions with noise of 0.5 pixels, but the number of outliers was gradually increased. As expected, the limiting factor is the Monte-Carlo sampling. As the inlier fraction decreases, more and more samples are needed to obtain any correct candidates for the selection process. When 75 outliers ($\approx 40\%$) are reached, which do not belong to any motion, the method gradually breaks down. It can be seen, from the estimated standard deviations and inlier numbers, that more outliers do not seriously impair scale estimation and model selection. Motions are simply missed, if no correct candidate is generated during sampling. In accordance with the theory, fundamental matrices are missed more often because of the larger required minimum sample. The experiment was also repeated with a higher sample number of 25,000/6,250. The results are slightly better, but on the whole, they confirm that the method can cope well with outliers, as long as the number of outliers is not significantly larger than the number of inliers *per model*. The results are summarized in Fig. 5.

5.2 Experiments with Real Image Pairs

We have also tested the proposed method on real image pairs. The first example contains three independently moving objects. On each of the three regions, 50 correspondences were measured manually, in order to have an easily accessible ground truth segmentation. Fifty spurious

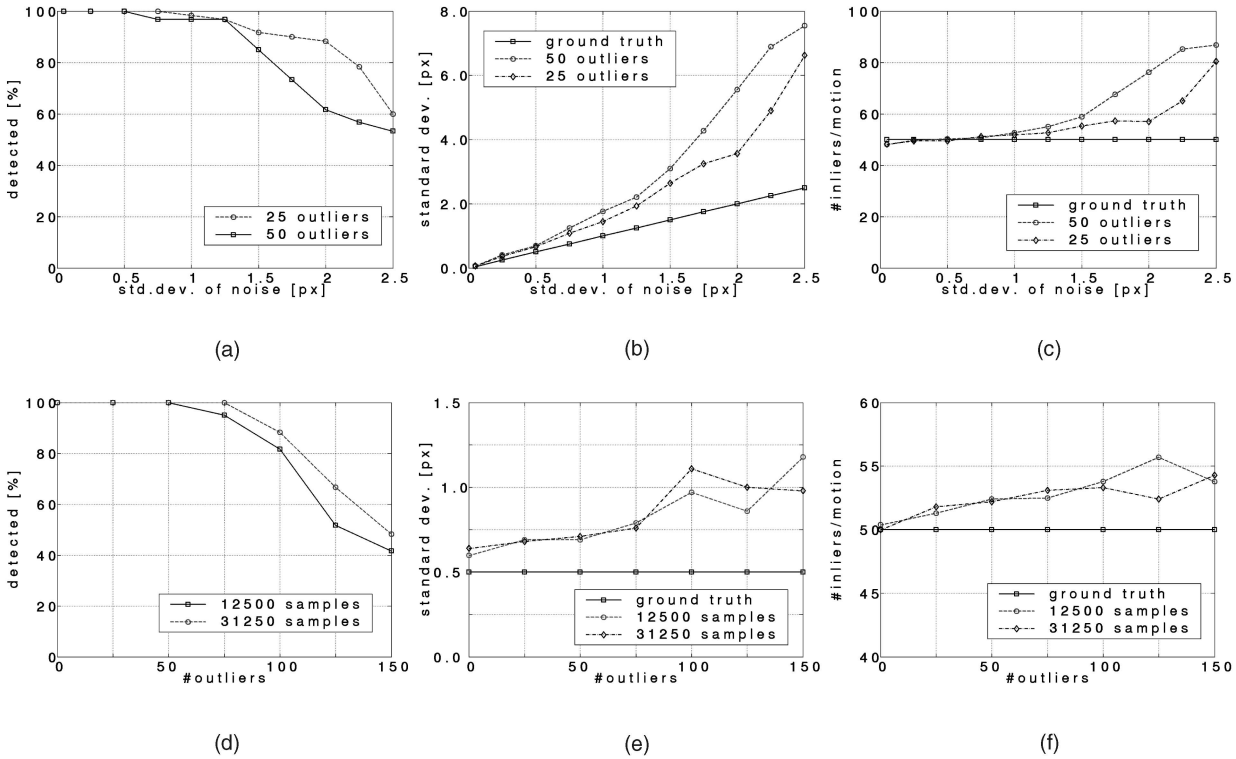


Fig. 5. Three-dimensional segmentation with synthetic data. Top row: Results at different noise levels. Bottom row: Results with different amount of outliers. See text for details. (a) estimated #motions, (b) estimated noise, (c) estimated #inliers, (d) estimated #motions, (e) estimated noise, and (f) estimated #inliers.

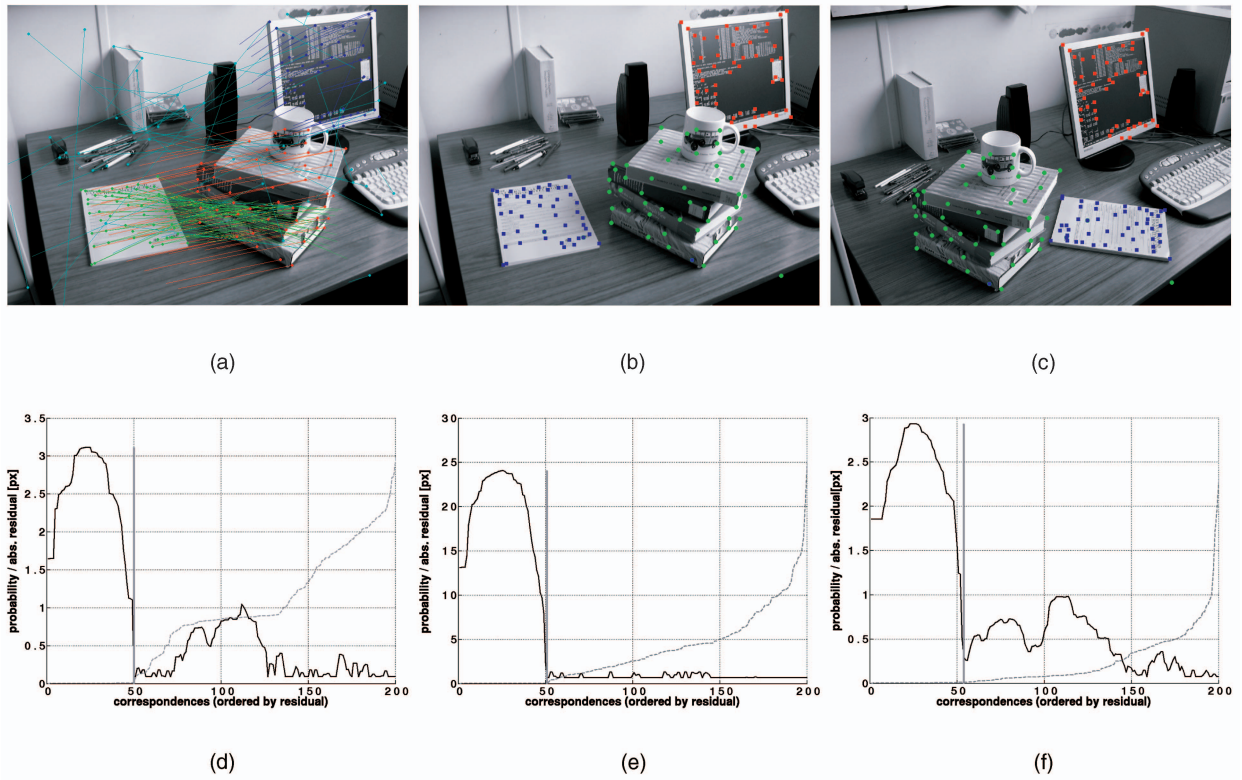


Fig. 6. Three-dimensional segmentation results for "desk" image pair. (a) Disparities of corresponding points overlayed on left image. Different colors denote different motions, cyan are outliers. (b) and (c) Obtained segmentation overlayed on images. Circles denote points on fundamental matrices, squares are points on homographies. Point classified as outliers are not displayed. (d), (e), and (f) Absolute residuals (gray, dashed), probability density function (black, continuous), and separation between inliers and outliers for the selected motion models. Probability densities are on a relative scale and do not integrate to 1. (a) Disparities, (b) left image, (c) right image, (d) pdf for screen motion, (e) pdf for journal motion, and (f) pdf for books motion.

matches were added at apparent intersections, repetitive structures, etc. Of 8,000 initial candidates, 17 fundamental matrices and 33 homographies survived the constraint ($\sigma_i < 4$ pixels) and were passed on to the model selection stage, which correctly retained one fundamental matrix for the pile of books and two homographies for the screen and the journal. Table 2 shows the obtained clustering of the matches. Ninety-nine percent of all inliers were assigned to the correct motion.

We have not disambiguated points which satisfy more than one motion model. A common strategy is to assign each point to the motion where it has the smaller (normalized) residual and, thus, the higher likelihood. However, this is theoretically questionable: The point is an inlier to both distributions and other information is necessary if it has to be disambiguated. Arguably, it is better (and closer to the human visual system) to assign it to the motion model satisfied by most of its neighbors.

TABLE 2
Three-Dimensional Segmentation Results for "Desk" Image Pair

object	motion	true	inliers	corr. inliers
books	F	50	53	50
journal	H	50	50	49
screen	H	50	49	49
outliers	—	50	*51*	*49*

The outliers are a rejection class for points not assigned to any motion. See text for details.

To demonstrate the importance of properly treating model overlap, the experiment has been repeated without correcting the joint likelihood of overlapping models (i.e., $q_{ij} = 0$ for all $i \neq j$). The result is a data description with eight fundamental matrices and 33 homographies, including the correct ones.

For the next experiment, an image pair was recorded and interest points were obtained automatically using the MSER detector. Each region was approximately normalized by diagonalizing its covariance matrix and removing the scale anisotropy, then the regions were matched with normalized cross-correlation, yielding 307 matches, including 50 outliers. The centers of gravity of each matching pair were used as correspondences. The proposed method correctly detected two fundamental matrices and one homography for the three moving objects in the scene. Ninety-six percent of the matches were assigned to the correct motion.

Further experiments where carried out with real images. The "shoes" sequence consists of 40 frames showing the feet of a moving person. There are three motions for the left foot, the right foot, and the (static) background. With the KLT-tracker, 72 points were automatically detected and tracked through the sequence, including two outliers due to tracking errors. The method was applied to the first and last image of the sequence and correctly segmented the data into three motions. A fundamental matrix was selected for the right shoe, and homographies were selected for the left shoe (where the tracked points are almost coplanar) and for the background

Fig. 7. Three-dimensional segmentation results for "box-book-mag," "shoes," and "truck" image pairs. Left: Disparities of corresponding points overlayed on left image. Different colors denote the ground truth segmentation, cyan are outliers. Center, right: Segmentation overlayed on images. Circles denote points on fundamental matrices, squares are points on homographies.

(which did not move at all). Ninety-four percent of the points were assigned to the correct motion. The two outliers could not be found since both accidentally satisfy the fundamental matrix. Again, additional information, such as spatial consistency, would be required to detect these cases [27].

Another image sequence was taken from the digitized 1946 movie "Transportation" (available from http://www.archive.org). The sequence contains two motions, the truck driving along the road and the background, which has a small motion due to camera jitter. Three hundred and five points were tracked through the sequence with the KLT-tracker, of which 32 are outliers. The method correctly recovered a fundamental matrix for the truck and a homography for the background motion; 97 percent of the points where assigned to the correct motion. Results for these experiments are depicted in Fig. 7.

5.3 Nonuniform Priors

To demonstrate the effect of the prior given at the end of Section 3, we have applied our method to the first and last image of the "car-truck-box" sequence also used by Vidal et al. [39], [40]. The data set contains three different motions with 44, 48, and 81 matches, respectively. Two of the motions are small and have ambiguous interpretations. Theoretically, both the car and the truck are nonplanar objects with general motion. However, the average Sampson residual when fitting a fundamental matrix to the matches on *the car and the truck together* is only $s_{F,ct} = 0.15$ pixels, while the average Sampson residual for the box is $s_{F,b} = 0.53$ pixels. Moreover, the two motions are so small that the average Sampson error for fitting homographies is $s_{H,c} = 0.13$ pixels for the car and $s_{H,t} = 0.44$ pixels for the truck, compared to $s_{F,c} = 0.07$ and $s_{F,t} = 0.11$ for fundamental matrices.

Fifty outliers were added by sampling spurious matches from a uniform distribution. Then, the method was applied to the data, using the prior from (24) with different values for U. The results are depicted in Fig. 8. With a uniform prior $U = 1$, two fundamental matrices are recovered: one for the box, and one for the truck and car *together*, since, even so, the fitting error is lower than for the box due to the degenerate configuration. With $U = [5 \ldots 6]$, the motions of the car and the truck are separated and assigned two homographies. With $U = [7 \ldots 12]$, the truck is assigned a fundamental matrix instead and, with $U = 13$, each motion is modeled by a fundamental matrix. Decreasing the model cost even further produces spurious models. The example illustrates nicely that there are multiple plausible interpretations of the same data, and a model selection criterion cannot be designed generically, but only for a certain task.

6 Concluding Remarks

We have presented a scheme for robust multibody structure-and-motion in the presence of different motion

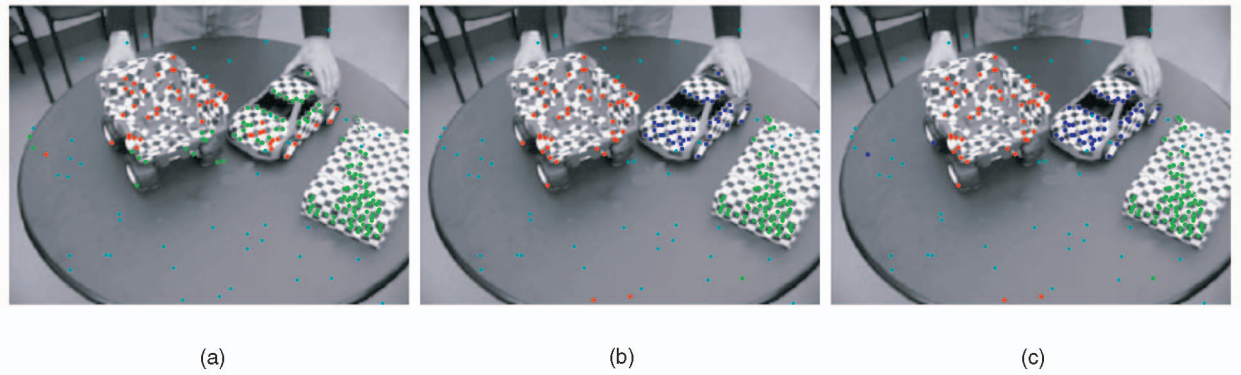


Fig. 8. Three-dimensional segmentation of "cars" image pair. Left image and segmentation results with priors of different strength. Colors denote the obtained clusters, circles denote points on fundamental matrices, squares are points on homographies. Cyan points are outliers. Left to right: uniform prior ($U = 1$), weak prior ($U = 7$), strong prior ($U = 13$). Details are given in the text.

models, noise of unknown standard deviation, and outliers. The method simultaneously recovers all present motions and needs no thresholds, except for an upper bound of the allowable measurement error. An important limitation is that the method is based on a set of candidate motions generated with random sampling. It therefore relies on a heuristic local scheme to keep the number of required samples in a manageable order of magnitude. Even so, the method can handle only a small number of motions.

A further requirement is that the numbers of correspondences on different motions are of the same order of magnitude. This restriction is a direct and inevitable consequence of the fact that the method is robust to outliers. An outlier model is effectively a rejection class for points not compatible with any motion model. Since the number of motions needed to explain the data is unknown, there has to be some complexity penalty to avoid overfitting (even if this penalty is not imposed explicitly as complexity cost, but implicitly, for example, in the form of a clustering threshold). Hence, if the number of points supporting some motion model $\mathcal{M}_S$ is only a small fraction of the total number of points, which includes the support for the larger motion $\mathcal{M}_L$, there must come a stage where it is cheaper to assign those points to the outliers than to add $\mathcal{M}_S$ to the model. In the presence of nonzero measurement noise, allowing for outliers inherently introduces a limit for the smallest identifiable motion.

The ideas underlying the presented method, including the limitations, are generic for robustly fitting multiple manifolds and not limited to structure-and-motion. In fact, among the potential applications, multibody structure-and-motion is on the challenging end of the scale because of the need to fit 3D manifolds and to decide between manifolds of varying dimension.

Finally, we reiterate that the task to be solved determines what a suitable model is. For example, we have shown that a compact description with sufficiently small errors on one hand and a fine-grained detection of all separable motions on the other hand requires different priors. In more general terms: A model selection problem cannot be solved with a generic criterion which is independent of the task. Rather, it has been shown that different model selection criteria are members of a larger family of priors on the model complexity and that, within this family, a problem-specific criterion can be designed, which incorporates all prior knowledge. The design and use of such priors has been briefly discussed in an ad-hoc manner, but further research is needed to establish a theoretically sound way of designing or learning priors for geometric model selection.

Acknowledgments

The authors would like to thank Hanzi Wang for help with the TSSE-estimator, Horst Bischof and Ales Leonardis for Taboo-search code, Jiri Matas for the MSER detector, and Rene Vidal for providing the "car-truck-box" data. They are also grateful to the anonymous reviewers for constructive comments which helped to improve the paper. This work has been carried out within the Institute for Vision Systems Engineering funded by the Faculty of Engineering, Monash University.

References

[1] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," *Proc. Second Int'l Symp. Information Theory,* pp. 267-281, 1973.

[2] S. Avidan and A. Shashua, "Trajectory Triangulation: 3D Reconstruction of Moving Points from a Monocular Image Sequence," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 4, pp. 348-357, Apr. 2000.

[3] G.L. Bretthorst, "An Introduction to Model Selection Using Probability Theory as Logic," *Maximum Entropy and Bayesian Methods,* G.R. Heidbreder, ed., pp. 1-42, Kluwer Academic Publishers, 1996.

[4] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 603-619, May 2002.

[5] O. Faugeras, Q.-T. Luong, and T. Papadopoulo, *The Geometry of Multiple Images.* MIT Press, 2001.

[6] O.D. Faugeras, "What Can Be Seen in 3D with an Uncalibrated Stereo Rig?" *Proc. Second European Conf. Computer Vision,* pp. 563-578, 1992.

[7] D.A. Forsyth and J. Ponce, *Computer Vision—A Modern Approach.* Prentice Hall, Inc., 2003.

[8] E.I. George and D.P. Foster, "Calibration and Empirical Bayes Variable Selection," *Biometrika,* vol. 87, no. 4, pp. 731-747, 2000.

[9] F. Glover and M. Laguna, "Tabu Search," *Modern Heuristic Techniques for Combinatorial Problems,* 1993.

[10] M. Han and T. Kanade, "Reconstruction of Scenes with Multiple Linearly Moving Objects," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 542-549, 2000.

[11] R. Hartley, "Estimation of Relative Camera Positions for Uncalibrated Cameras," *Proc. Second European Conf. Computer Vision,* pp. 579-587, 1992.

[12] R. Hartley, "Projective Reconstruction and Invariants from Multiple Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, pp. 1036-1041, 1994.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge Univ. Press, 2000.
[14] K. Huang, Y. Ma, and R. Vidal, "Minimum Effective Dimension for Mixtures of Subspaces: A Robust GPCA Algorithm and Its Applications," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 631-638, 2004.
[15] P.J. Huber, *Robust Statistics.* John Wiley and Sons, 1981.
[16] M. Irani and P. Anandan, "A Unified Approach to Moving Object Detection in 2D and 3D Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 6, pp. 577-589, June 1998.
[17] K. Kanatani, *Statistical Optimation for Geometric Computation: Theorie and Practice.* Elsevier, 1996.
[18] K. Kanatani, "Geometric Information Criterion for Model Selection," *Int'l J. Computer Vision,* vol. 26, no. 3, pp. 171-189, 1998.
[19] K. Kanatani, "Uncertainty Modeling and Model Selection for Geometric Inference," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 10, pp. 1307-1319, 2004.
[20] A. Leonardis, A. Gupta, and R. Bajcsy, "Segmentation of Range Images as the Search for Geometric Parametric Models," *Int'l J. Computer Vision,* vol. 14, no. 1, pp. 253-277, 1995.
[21] C. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene from Two Projections," *Nature,* vol. 293, pp. 133-135, 1981.
[22] Y. Ma, J. Kosecka, S. Soatto, and S. Sastry, *An Invitation to 3-D Vision.* Springer Verlag, 2003.
[23] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," *Proc. 13th British Machine Vision Conf.,* vol. 1, pp. 384-393, 2002.
[24] C. Matsunaga and K. Kanatani, "Calibration of a Moving Camera Using a Planer Pattern: Optimal Computation, Reliability Evaluation and Stabilization by Model Selection," *Proc. Sixth European Conf. Computer Vision,* vol. 2, 2000.
[25] J. Rissanen, "Modeling by Shortest Data Description," *Automatica,* vol. 14, pp. 465-471, 1978.
[26] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection.* John Wiley and Sons, 1987.
[27] K. Schindler, "Spatially Consistent 3D Motion Segmentation," *Proc. Int'l IEEE Conf. Image Processing,* 2005.
[28] G. Schwartz, "Estimating the Dimension of a Model," *Annals of Statistics,* vol. 6, pp. 497-511, 1978.
[29] C.E. Shannon, "A Mathematical Theory of Communication," *Bell Systems Technical J.,* vol. 27, pp. 379-423, 1948.
[30] A. Shashua and A. Levin, "Multiframe Infinitesimal Motion Model for the Reconstruction of (Dynamic) Scenes with Multiple Linearly Moving Objects," *Proc. Eighth Int'l Conf. Computer Vision,* pp. 592-599, 2001.
[31] M. Stricker and A. Leonardis, "ExSel++: A General Framework to Extract Parametric Models," *Proc. Computer Analysis of Images and Patterns,* pp. 90-97, 1995.
[32] P. Sturm, "Structure and Motion of Dynamic Scenes—The Case of Points Moving in Planes," *Proc. Seventh European Conf. Computer Vision,* pp. 867-882, 2002.
[33] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," Technical Report CMU-CS-91-132, Carnegie Mellon Univ., 1991.
[34] W.-S. Tong, C.-K. Tang, and G. Medioni, "Simultaneous Two-View Epipolar Geometry Estimation and Motion Segmentation by 4D Tensor Voting," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1167-1184, Sept. 2004.
[35] P.H.S. Torr, "Geometric Motion Segmentation and Model Selection," *Philosophical Trans. Royal Soc. London A,* vol. 356, no. 1740, pp. 1321-1340, 1998.
[36] P.H.S. Torr, "Model Selection for Structure and Motion Recovery from Multiple Images," Technical Report MSR-TR-99-16, Microsoft Research, 1999.
[37] P.H.S. Torr, "Bayesian Model Estimation and Selection for Epipolar Geometry and Generic Manifold Fitting," *Int'l J. Computer Vision,* vol. 50, no. 1, pp. 35-61, 2002.
[38] R. Vidal and Y. Ma, "A Unified Algebraic Approach to 2-D and 3-D Motion Segmentation," *Proc. Eighth European Conf. Computer Vision,* pp. 1-15, 2004.
[39] R. Vidal and S. Sastry, "Optimal Segmentation of Dynamic Scenes from Two Perspective Views," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003.
[40] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "Segmentation of Dynamic Scenes from the Multibody Fundamental Matrix," *Proc. ECCV Workshop Visual Modeling of Dynamic Scenes,* 2002.
[41] C.S. Wallace and D.M. Boulton, "An Information Measure for Classification," *Computer J.,* vol. 11, no. 2, pp. 185-194, 1968.
[42] M.P. Wand and M. Jones, *Kernel Smoothing.* Chapman and Hall, 1995.
[43] H. Wang and D. Suter, "MDPE: A Very Robust Estimator for Model Fitting and Range Image Segmentation," *Int'l J. Computer Vision,* vol. 59, no. 2, pp. 139-166, 2004.
[44] H. Wang and D. Suter, "Robust Fitting by Adaptive-Scale Residual Consensus," *Proc. Eighth European Conf. Computer Vision,* pp. 107-118, 2004.
[45] L. Wolf and A. Shashua, "Two-Body Segmentation from Two Perspective Views," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 263-270, 2001.
[46] Z. Zhang, "Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting," *Image and Vision Computing,* vol. 15, no. 1, pp. 59-76, 1997.

Konrad Schindler received the Diplomingenieur degree in photogrammetry from the Vienna University of Technology, Austria, in 1999, and the PhD degree from the Graz University of Technology, Austria, in 2003. He worked as a photogrammetric engineer in private industry and was a research assistant in the Computer Graphics and Vision Department of the Graz University of Technology. He is currently a postdoctoral research assistant in the Digital Perception Lab of Monash University in Melbourne, Australia. His research interests include the analysis and reconstruction of dynamic scenes, as well as feature detection and object recognition. He is a member of the IEEE.

David Suter received the BSc degree in applied mathematics and physics from Flinders University, Adelaide, Australia, in 1977 and the PhD degree in computer vision from LaTrobe University, Melbourne, Australia, in 1991. He is an associate professor in the Department of Electrical and Computer Systems Engineering at Monash University, Melbourne, Australia. He served as general cochair of the 2002 Asian Conference on Computer Vision and has been cochair of the Statistical Methods in Computer Vision workshops (2002 Copenhagen and 2004 Prague). He currently serves on the editorial board of the *International Journal of Computer Vision* and of the *International Journal of Image and Graphics*. His main research interest is motion estimation from images and visual reconstruction. He is a senior member of the IEEE and the IEEE Computer Society and vice president of the Australian Pattern Recognition Society.