Department of Electrical and Computer Systems Engineering

Technical Report MECSE-12-2004

Simultaneous, robust fitting of multiple 3D motion models

K. Schindler

MONASH UNIVERSITY

Simultaneous, robust fitting of multiple 3D motion models

Konrad Schindler
Insititute for Vision Systems Engineering, Monash University
Wellington Road, Clayton, 3800 Victoria, Australia

October 26, 2004

Abstract

Multi-body 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 is not static, but consists of multiple rigid object moving relative to each other. In this report the case of two views is examined. The general 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. The correspondences can be constrained by either a fundamental matrix (non-planar object, general motion) or a homography (planar object or pure rotation). 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 to explain the measurements. The problem is a special case of the general problem of robust model selection and fitting. The proposed solution follows the recover-and-select paradigm: it randomly creates a large number of candidate models, estimates the scale of the noise for each model, and computes its goodness-of-fit. Then model selection is used to prune the redundant collection of models to an optimal set, including an outlier model.

1 Introduction

In the last decade, structure-and-motion recovery from perspective images as the only source of data 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 [17], Faugeras [6] and Hartley [10] to the comprehensive theory now presented in several excellent textbooks [12, 5, 18]. Among other results, it turned out that not all scenes and not all relative camera positions can be described by the most general motion model, the epipolar geometry, encoded algebraically by the *fundamental matrix*. There are two cases, in which the fundamental matrix becomes degenerate and must be replaced by a more restrictive model [5]. The first one is due to a special motion: if the camera only rotates around its projection center, 3D-reconstruction is obviously not possible, because the rays through a scene point are identical and no triangulation is possible. Instead, the two sets of image points are related by a projectivity, algebraically expressed as a *homography*, a projective mapping from one image plane to the other one. The second degenerate configuration is due to a special scene structure: if the viewed scene is planar, the two image planes are also related by a projectivity (as each of them is related to the scene plane by a homography). In the following, we will assume that the effects of perspective projection are noticeable and only consider these two motion models, however the framework is completely general and can be extended to other, simpler models, as for example shown in [27].

To select the correct motion model for a scene, it is necessary to compare the goodness-of-fit of different models with a suitable model selection criterion, which finds a balance between the fitting error and the dimension and complexity of the model. The first applications of model selection to selecting a two-view motion model are due to Kanatani [15] and to Maybank and Sturm [19]. The

fact that the dimension of the fitted manifold requires separate treatment was first recognized in the computer vision literature by Kanatani [14].

Soon after the main SaM-theory had been established, researchers turned to the more challenging case of *dynamic* scenes, but the geometric properties of dynamic scenes turned out to be non-trivial even for simple motions [2, 9, 24, 34, 26]. Recently an excellent extension of algebraic SaM-theory to dynamic scenes has been presented [31, 30, 11]. However 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"), and is not designed to use other motion models. Specifically, it does not allow for homographies and it does not include an outlier model. The latter, together with the non-linear nature of the problem, makes the multibody fundamental matrix potentially vulnerable to gross measurement errors.

A different way to tackle the problem is not to extend the geometric model, but instead try to cluster the points according to their motion. However 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. Therefore, previous approaches have adopted an iterative strategy: a single motion is estimated, the points consistent with it are searched and 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. Therefore, only a preliminary result is obtained, which has to be post-processed with expectation maximization and pruning of motion models [27].

As already said earlier, also in the context of MSaM the need arises to select the correct motion model for each motion. In the iterative approach, the models are disjoint, and the likelihood of all used motion models can be directly added up to gain a new model selection criterion. This extension has also first been studied by Torr [27].

We will look at the two-view MSaM-problem (a generalization of Torr's method to 3 views is given in [20]). The question we want to answer is: Which is the optimal collection of (an unknown number of) fundamental matrices and homographies to explain the image correspondences, allowing for outliers, which do not belong to any of these models, and allowing for unknown (and possibly varying) standard deviation of the models? Our view on the problem is the one of robust statistics and model selection. The rationale is the following: if we had a set of motion models including the correct ones, it would be easier to tell the ones we need to explain the data from the redundant ones. Hence, we follow a recover-and-select scheme. In a first step, motion models are instantiated by Monte-Carlo sampling from the observed correspondences. Non-parametric statistical analysis of the residuals is used to robustly estimate the scale of the noise for each model in spite of the low number of inliers. Given the scale and the number of inliers found with this scale, the likelihood of each model can be computed. The likelihood then is used as a measure for the goodness-of-fit to select an optimal set of motion models with a model selection criterion.

There are two original contributions in this paper, one in each step. Firstly, other than previous approaches, the presented method is able to estimate the scale of the noise from the data. Compared with a globally preset threshold, this improves the capability to discriminate between different tentative models: a global threshold for inlier/outlier separation selects different arbitrary portions of the data with the same maximum residual, without taking into account the shape of the underlying distributions, and therewith obscures the statistical properties of the data: if the threshold is wider than the distribution and outliers are included, then the number of inliers and the fitting residuals are over-estimated; if the threshold is too narrow, the inlier distribution is truncated and the two quantities are under-estimated. The incorrect estimates for the threshold, inlier number and fitting residual will influence the subsequent model selection, because these quantities are exactly the variables used to assess the goodness-of-fit. In contrast, the new method explicitly recovers the distribution of the residuals and for each tentative model estimates an individual standard deviation and separation between inliers and outliers.

Secondly, previous approaches to outlier-tolerant MSaM are iterative: they recover one motion model at a time, remove its inliers from the data, then search for the next motion in the remaining data. Motion models are thus regarded as statistically independent, which is clearly not true, since they may overlap (i.e., there are points which satisfy more than one model). In the presence of overlapping models, an iterative approach will assign such points to the model detected first, rather

then to the one they are most likely to belong to. Also, a model $\mathcal{A}$ may be a likely explanation, if a second model $\mathcal{B}$ is *not* used (because $\mathcal{A}$ also explains most of $\mathcal{B}$'s points), but not if $\mathcal{B}$ is used (because $\mathcal{B}$ better explains the overlap and the remainder of $\mathcal{A}$ has large residuals). This paper demonstrates non-iterative, simultaneous selection of all models. A new formulation for the posterior likelihood is derived, which properly treats the overlap between models. Maximizing the posterior can then be posed as quadratic boolean optimization, so that the joint likelihood between models is taken into account. Selecting a set of models and finding their respective inliers becomes a one-shot procedure and does not require a subsequent EM-step, which has notoriously poor convergence.

2 Generating candidate models

For subsequent model selection, a set of candidate models has to be generated. This is done with a simple Monte-Carlo procedure: models are randomly instantiated from a minimal set of correspondences (7 for a fundamental matrix, 4 for a homography). Unfortunately, the presence of multiple motions means that only a comparatively small fraction of all correspondences belongs to each model. Therefore a localized sampling scheme is required to ensure that good candidates are found.

2.1 Localized sampling

A critical part of the proposed scheme is to ensure that good candidates are found for all present motions. If candidates for one or more motions are missing in the candidate set, they can obviously not be added at a later stage. Worse still, model selection could potentially be mislead by the attempt to find an explanation for the data points on missed models and give wrong results. Applying brute-force random sampling is already very expensive, if 2 motions are present, and becomes intractable for more than 2 motions: if we assume that the smallest inlier set comprises 20% of the data (an optimistic guess for 3 motions and some outliers), the standard formula for RANSAC shows that we would need $\frac{\log(0.99)}{\log(1-0.2^7)} = 359777$ samples. Even if completely awkward samples are discarded at an early stage, all tentative models have to be analyzed with mean-shift at least once (see next section), so that this figure is an order of magnitude too high for practical applications.

A practical solution to the problem is to exploit the spatial coherence of points belonging to the same motion. Except for special cases such as transparent objects, the points belonging to the same rigid object will normally be clustered in the image plane. In a certain region of the image plane, the points will belong predominantly to one of the present motions, and a local sampling scheme will therefore dramatically reduce the number of samples required to find an uncontaminated one. To achieve computational efficiency, it is crucial to carefully design a local sampling scheme, e.g. sample from many small subsets, if several small moving objects are observed, or sample from cells of different pyramid levels, if objects of very different size are observed.

For the experiments in section 4, the image plane was subdivided into 3 overlapping rows and 3 overlapping columns, and samples were drawn from the entire image, each column, each row, and each of the 9 regions defined by a row-column intersection (see Figure 1). This heuristic subdivision scheme proved to be efficient for different images. It is a compromise between local coherency and global extension. To justify the plausibility of the heuristics, we may say the following: On one hand, if a large object is present, it will cover a large image area (possibly all of the image). In this case, one should sample from this large area, in order to obtain well distributed points for the estimate. However, if there are not too many outliers, a moderate number of samples will be sufficient, because there are no points in the area belonging to other objects. On the other hand one column-row intersection in the scheme covers 11% of the entire image plane, the overlap with the neighboring column-row intersection covers 5.5%. Let us assume that an independently moving object covers at least 10% of the image (smaller objects will in general give poor motion estimates), and that it is not very elongated in shape. Then there is at least one region, in which

the object covers ≈50% of the entire area, which (except for the outliers) again requires < 600 samples per region.

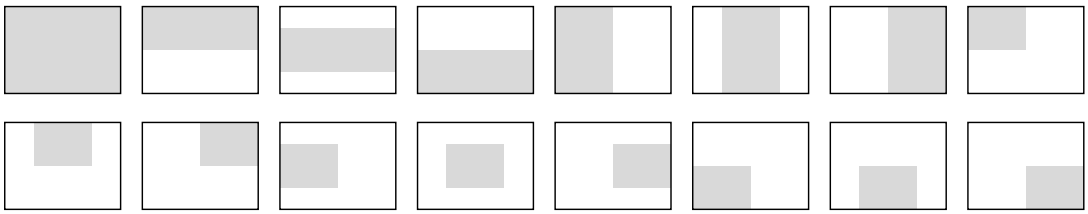


Figure 1: Localized sampling scheme for tentative motion models. Samples are drawn from subregions of the image plane to exploit spatial coherence and reduce the required sample number.

2.2 Estimating the scale of the noise

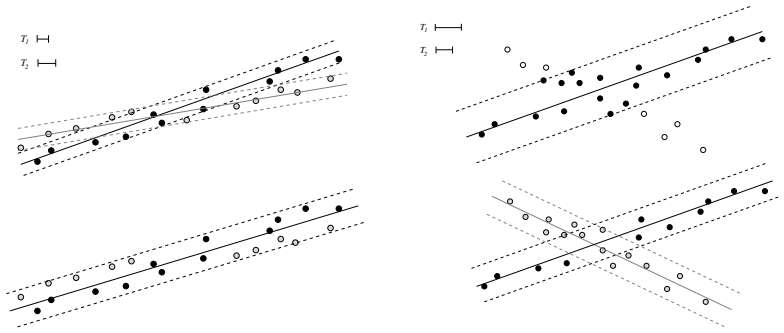
Given a model and a number of data points, the scale of the noise can be estimated without any further knowledge by applying the MDPE-estimator of Wang and Suter [32]. To this end, the residuals of all data points w.r.t. the model have to be computed. To quantify the residual of each point w.r.t. a fundamental matrix or homography the linear approximation of the geometric error or "Sampson-error" is used [12]. Given the ordered set of residuals for a model, the mean-shift method [4] is used to obtain a nonparametric estimate for their probability density. Assuming that the inliers have mean zero, the valley of this density function, which is closest to 0, is a sensible point to separate inliers from outliers [32]. The process is illustrated in Figure 5(c)-5(e): for three different models, the ordered absolute residuals and the corresponding probability density functions are superimposed. The vertical black bars indicate the estimated boundary between inliers and outliers.

In a two-step process, the bandwidth for the mean-shift algorithm can be selected automatically from the data with an oversmoothed bandwidth-selector [33]. With the approximate inlier set, standard methods from robust statistics yield an estimate for the standard deviation.

Estimating the variance and inlier threshold of each model separately from the data considerably improves the power of the method, compared with a RANSAC-like method with a fixed threshold between inliers and outliers. When searching for a *single* model, a slightly incorrect threshold is not problematic, while it may impair the results in the presence of multiple models. There are two possible cases: if the threshold is too low, not all inliers are found, however the model 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 models, the situation is different. If only a subset of the inliers is found and assigned to the model, the remaining inliers will give rise to a second, similar model, leading to overfitting (see Figure 2(a)). 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 model a robust least-squares technique such as an M-estimator [13] can be used to obtain a correct fit, as suggested for example by Triggs [29]. Again, the situation is more complicated in the presence of multiple models: if two models overlap, the overlap will be overestimated with the larger threshold. This can lead to underfitting, because one of the models (either the one with larger support or simply the one detected first) will claim too many of the data points, leaving only little support for the second model (see Figure 2(b)).

2.3 Refinement

As noted by Rousseeuw [22] and confirmed by other authors, e.g. Zhang [35], the *efficiency* of random sampling methods is poor, i.e., even a model constructed from the best uncontaminated random sample may differ quite strongly from the optimal fit. Therefore, it is necessary to refine each tentative model with a least-squares fit to the inlier points. To avoid unnecessary computations, it is advisable to discard candidate models with very low number of inliers before the



(a) Low threshold (b) High threshold

Figure 2: Influence of wrong thresholds on fitting multiple models. (a) Too low thresholds encourage overfitting: data points missed by the fit (black line) give rise to another model with low residuals (gray line). (b) Too large thresholds encourage underfitting: data points wrongly assigned to the fit (black line) weaken the support for other models (gray line).

expensive refinement step. One could derive the minimum number of inliers for a model from (11) by demanding $q_{ii} > 0$. For a conservatively chosen range of realistic parameters

- $\sigma_i = 0.05..5$ pixels
- $N_t = 50..5000$ features
- $A = 0.25 \ldots 16$ MPixel

this yields 4-8% of the total number of features N_t for a fundamental matrix and 2-5% for a homography. Practically, we can discard features with $N_i < \frac{1}{10} N_t$, since it will not be possible to find models with lower inlier fraction by random sampling. The threshold should be chosen such that it only discards useless samples, but does not influence the selection process!

3 Model Selection

3.1 Principle of Geometric Model Selection

To select the optimal set of models, a model selection criterion is needed, which balances the goodness-of-fit against the complexity of the complete description by penalizing the addition of new motion models depending on their dimension and cardinality. There are several model selection criteria in the statistical literature, starting with Akaike's *an information criterion* AIC [1]. Although his pioneering work introduced the basic principle which was then refined in most other model selection methods, 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. Standard model selection criteria, which remedy this problem, are Schwartz' *Bayes information criterion* BIC [23], an approximation of the a-posteriori likelihood, and Rissanens *minimum description length* MDL [21], an information theoretic criterion that seeks to minimize the coding length of the data. In spite of a completely different derivation, the two surprisingly yield similar criteria.

However, all these criteria in their standard form assume that the dimension of the fitted model is known and only the number of parameters of that model varies. Since we have to decide between models of different dimension, an extension is needed, which penalizes not only the number of parameters, but also the dimension – otherwise the model 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 smaller or equal the errors w.r.t. 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 [14]. However, it has already been stated that AIC has theoretical weaknesses. A similar extension for BIC (or MDL), based on Bayesian decision theory, is the core of Torr's work on selecting motion models, which is the closest work to the one presented here. The criterion he proposes is termed *geometrically robust information criterion* GRIC [27, 28]. We share Torr's view that the theoretical foundation of BIC/MDL is stronger than the one of AIC – although we are aware of the fact that there is no "objective, canonical" way to select a model and that it is therefore useless to argue why one criterion should be "more correct" than another one.

We will use GRIC in the rest of this 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. GRIC selects the model $\mathcal{M}$ which maximizes

$$\mathrm{GRIC}(\mathcal{M}) = 2\ln(\mathcal{L}) - N_t D \ln(R) - K \ln(RN_t) \tag{1}$$

where N_t is the total number of correspondences, R is the dimension of the data (4 for pairs of image points), K is the number of parameters of the motion model (8 for a homography, 7 for a fundamental matrix), and D is the dimension of the manifold (2 for a homography and 3 for a fundamental matrix). $\mathcal{L}$ is the likelihood of the model.

Note the difference between the number of required points and the number of free parameters to estimate a model: for a fundamental matrix $U_i = K_i = 7$, while for a homography $U_i = 4, K_i = 8$ because of the different dimension. To compute the 7 parameters of a fundamental matrix 7 correspondences are needed. However, to compute the 8 parameters of a homography, only 4 points are needed, as each point gives 2 equations. Therefore the variance estimated from the Sampson residuals is $\sigma^2 = \frac{1}{N-4}\sum \epsilon^2$. The variance *per coordinate* is a factor of 2 lower: $\sigma^2_{x,y} = \frac{1}{2N-8}\sum(\epsilon_x^2 + \epsilon_y^2) = \frac{1}{2}\frac{1}{N-4}\sum \epsilon^2$.

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. Following the Bayesian view advocated by Bretthorst [3], among others, we choose the least informative distributions, where the Shannon entropy is used as a measure of how informative a distribution is: assuming that the inlier distribution is symmetric with zero-mean, the least informative one is a Gaussian, while a uniform distribution within the image boundaries is the least informative distribution for the outliers (given no further information).

Since we want to select a subset of the set of models established previously, the likelihood has to be split into the contributions each single model makes to the total likelihood. At the same time, we have to account for the fact that models may overlap, i.e., data points may be inliers t more than one model. The data points in the overlap should of course contribute only once to the overall likelihood (explaining a point more than once does not increase the likelihood). We will from now on assume that there is only pairwise overlap between the models. This assumption is not strictly correct and causes too large overlap penalties (in general there will be points which are inliers to more than 2 motion models), but the number of these points is small compared to the number of points which are inliers to more than one model. This approximation is necessary to yield a tractable optimization problem, as explained in more detail in section 3.3.

Let $\mathcal{V}_i$ denote a tentative motion model with standard deviation σ_i, and let $\{\mathbf{p}_k, k \in \mathcal{V}_i\}$ denote the N_i points, which are inliers to $\mathcal{V}_i$. Furthermore, let $\epsilon_{(i),k}$ be their residuals w.r.t. $\mathcal{V}_i$. Then the likelihood of $\mathcal{V}_i$ is

$$\mathcal{L}_i = \prod_{k \in \mathcal{V}_i} \left(\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{\epsilon^2_{(i),k}}{2\sigma_i^2} \right) \right) \tag{2}$$

Now let us introduce a second tentative model $\mathcal{V}_j$. If both models are used, and they overlap, then the overlapping points should contribute only to the likelihood of one of them, as there is

no benefit in "explaining the same point twice". Rather, each point should only contribute to the model, in which it has the higher likelihood. Let $\{\mathbf{p}_k, k \in \mathcal{V}_{[ij]}\}$ denote the $N_{[ij]}$ points, which are inliers to both models $\mathcal{V}_i$ 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 models were regarded as 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 model, 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 models is given by $\mathcal{L}_{i\cup j} = \frac{\mathcal{L}_i\mathcal{L}_j}{\mathcal{L}_{[ij]}}$, where

$$\begin{aligned}\mathcal{L}_{[ij]} &= \prod_{k\in\mathcal{V}_{[ij]}} \min\left(\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{\epsilon^2_{(i),k}}{2\sigma_i^2}\right),\ \frac{1}{\sigma_j\sqrt{2\pi}}\exp\left(-\frac{\epsilon^2_{(j),k}}{2\sigma_j^2}\right)\right) \\ &= \prod_{k\in\mathcal{V}_{[i]}}\left(\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{\epsilon^2_{(i),k}}{2\sigma_i^2}\right)\right)\prod_{k\in\mathcal{V}_{[j]}}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\exp\left(-\frac{\epsilon^2_{(j),k}}{2\sigma_j^2}\right)\right)\end{aligned} \tag{3}$$

Let the set of all candidate models be $\mathcal{C}$. If we denote a subset $\widehat{\mathcal{C}}$ of $\mathcal{C}$ by $\{\mathcal{V}_i, i \in \widehat{\mathcal{C}}\}$, and the likelihood of the outlier points w.r.t. $\widehat{\mathcal{C}}$ by $\mathcal{L}_{/\widehat{\mathcal{C}}}$, then the total likelihood of $\widehat{\mathcal{C}}$ is

$$\mathcal{L}_{\widehat{\mathcal{C}}} = \frac{\prod_{i\in\widehat{\mathcal{C}}}\mathcal{L}_i}{\prod_{i,j\in\widehat{\mathcal{C}}}\mathcal{L}_{[ij]}}\mathcal{L}_{/\widehat{\mathcal{C}}} \tag{4}$$

If no constraints are enforced when matching, then the probability density for matches which are outliers to all models is $P = \frac{1}{A^2}$, where A is the image area measured in the same units as σ, and the likelihood of N_p outliers is $\mathcal{L}_{/\widehat{\mathcal{C}}} = P^{N_p}$. The area A has to be changed appropriately, if the search area for matching is restricted, such as in a short-baseline setup or when tracking the points through a video sequence.

To compare different subsets $\widehat{\mathcal{C}}$, one can introduce a boolean vector $\mathbf{b}$, with elements $b_i = 1$ if model $\mathcal{V}_i$ is used, and $b_i = 0$ otherwise, so that $\widehat{\mathcal{C}} = f(\mathcal{C}, \beta)$. Then the total log-likelihood of the chosen subset is

$$\ln(\mathcal{L}) = \sum_{i\in\mathcal{C}}\left(b_i\ln(\mathcal{L}_i)\right) - \sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}}\left(b_ib_j\ln\left(\mathcal{L}_{[ij]}\right)\right) + N_p\ln(P) \tag{5}$$

In this expression we can substitute the likelihoods with expressions (2) and (3). 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) and obtain

$$\begin{aligned}\ln(\mathcal{L}) = &\left(N_t - b_i\sum_{i\in\mathcal{C}}N_i + b_ib_j\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}}N_{[ij]}\right)\ln(P) - \sum_{i\in\mathcal{C}}\left(b_i\left(\frac{N_i}{2}\ln(2\pi\sigma_i^2) + \frac{1}{2\sigma_i^2}\sum_{k\in\mathcal{V}_i}\epsilon^2_{(i),k}\right)\right) + \\ &+\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}}\left(b_ib_j\left(\frac{N_{[i]}}{2}\ln(2\pi\sigma_i^2) + \frac{1}{2\sigma_i^2}\sum_{k\in\mathcal{V}_{[i]}}\epsilon^2_{(i),k} + \frac{N_{[j]}}{2}\ln(2\pi\sigma_j^2) + \frac{1}{2\sigma_j^2}\sum_{k\in\mathcal{V}_{[j]}}\epsilon^2_{(j),k}\right)\right)\end{aligned} \tag{6}$$

To simplify equation (6), we drop the constant term $N_t\ln(P)$. This will not influence the maximization of the likelihood (for the sake of clarity, the new quantity is still called $\mathcal{L}$). Furthermore we abbreviate the normalized sum of squared errors of a model

$$\frac{1}{\sigma_i^2}\sum_{k\in\mathcal{V}_i}\epsilon^2_{(i),k} = N_i - U_i = E_i \tag{7}$$

and the normalized sums of squared errors in the overlap

$$\frac{1}{\sigma_i^2}\sum_{k\in\mathcal{V}_{[i]}}\epsilon_{(i),k}^2 = E_{[i]} \quad , \quad \frac{1}{\sigma_j^2}\sum_{k\in\mathcal{V}_{[j]}}\epsilon_{(j),k}^2 = E_{[j]} \tag{8}$$

Substituting these expressions in (6), setting $\lambda_1 = -2\ln(P) - \ln(2\pi)$, multiplying by 2 and re-ordering yields

$$\begin{aligned} 2\ln(\mathcal{L}) = &\sum_{i\in\mathcal{C}}\left(b_i\left(N_i\lambda_1 - N_i\ln(\sigma_i^2) - E_i\right)\right) \\ &- \sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}}\left(b_i b_j\left(N_{[ij]}\lambda_1 - N_{[i]}\ln(\sigma_i^2) - E_{[i]} - N_{[j]}\ln(\sigma_j^2) - E_{[j]}\right)\right) \end{aligned} \tag{9}$$

It is worth mentioning that scaling the images with a scalar β leaves the expression unchanged, $(\lambda_1 N_i - N_i\ln(\sigma_i^2) + N_i\ln(\beta^2) - N_i\ln(\beta^2) - \beta^{-2}\beta^2 E_i$, and analogous for the second term, proving that it is independent of the global scale, as required for a likelihood.

3.3 Maximizing the Criterion

Previously, model selection criteria where either used to select an unknown number of models with the same dimension at once, such as in [16], or to select one model of varying dimension at a time, as in [14, 27]. The machinery to solve the optimization problem, which is adopted in the following, stems from the former work, while the theory needed to cope with varying dimension stems from the latter. The additional constraint, that we have to formulate a tractable optimization problem for an unknown number of models, means that we have to separate the contributions of different models to the total likelihood, which is the reason that we assume only pairwise overlap: taking into account overlap between at most N models leads to an N-dimensional boolean optimization problem.

With expression (9) for the likelihood, the GRIC (1) for a model collection $\widehat{\mathcal{C}}(\mathbf{b})$ can be written as

$$\mathrm{GRIC}(\mathbf{b}) = \mathbf{b}^\mathsf{T}\mathtt{Q}\mathbf{b} \tag{10}$$

where $\mathtt{Q}$ is a symmetric matrix [16]. Let the constants $\lambda_2 = N_t\ln(4)$ and $\lambda_3 = \ln(4N_t)$. Then the diagonal elements of $\mathtt{Q}$ are

$$q_{ii} = N_i\lambda_1 - N_i\ln(\sigma_i^2) - E_i - \lambda_2 D_i - \lambda_3 K_i \tag{11}$$

and the off-diagonal elements are

$$q_{ij} = q_{ji} = -\frac{1}{2}\left(N_{[ij]}\lambda_1 - N_{[i]}\ln(\sigma_i^2) - E_{[i]} - N_{[j]}\ln(\sigma_j^2) - E_{[j]}\right) \tag{12}$$

Maximizing expression (10) over $\mathbf{b}$ is a combinatorial problem, for which a global optimum can only be found through exhaustive search, so we have to use a heuristic maximization technique (it seems, however, that for our problem the global maximum is always found). Taboo-search [8] is a standard method to solve this type of problems. In computer vision, it has been advocated, and applied for similar problems, by Stricker and Leonardis [25].

To round up this section, let us give an intuitive interpretation of equations (11) and (12): The cost function

- favors motions which explain a high number of data points (more precisely: which greatly reduce the number of outliers)
- favors models with low standard deviation

- tries to keep the model dimension low (i.e., a homography, which explains the same number of points with the same standard deviation is preferred over a fundamental matrix). This is the difference between GRIC and conventional model-selection criteria: it accounts for the varying dimension of the data, since a model with higher dimension *always* explains a data set as good as or better than a low-dimensional one, even if it is an overfit with poor predictive power.
- tries to keep the number of models low by penalizing each model proportional to the number of its parameters (the usual complexity penalty of model selection criteria)

Note that no arbitrary parameters have to be tuned in (11) and (12). If one accepts Bayesian decision theory as a basis and the Shannon-entropy as a measure of how informative a distribution is, the weights for the different terms of the selection criterion are determined.

3.4 Constraints

For any real problem the maximum allowable error ϵ_{max} for a single point measurement is known – it is the amount of error above which a measurement is considered an "outlier" rather than a "noisy inlier". In the presence of a single model without outliers, the maximum allowable error would be a natural upper bound for the standard deviation σ of a motion model, since $\frac{1}{N}\sum \epsilon_i^2 \leq \max(\epsilon_i^2)$. To account for outliers and pseudo-outliers on other motion models, which tend to blur the inlier/outlier boundaries, it is necessary to use a more conservative upper bound $t\epsilon_{max}, t \approx 2$.

However, the constraint that there is an upper bound for the Sampson error is not apparent in the objective function for model selection. During random sampling, it is possible (and in fact not unlikely in practical situations) that candidate models are found, which explain a large number of points originating from more than one motion model, albeit with overly high σ. The model selection formalism does not guarantee that their lower goodness-of-fit will compensate for the large number of explained points. To formally add the ϵ_{max}-constraint to the probabilistic formulation of the optimization problem, one would have to redefine the likelihood (2) of a candidate model $\mathcal{V}_i$ as

$$\mathcal{L}_i = \begin{cases} \prod_{k \in \mathcal{V}_i} \left(\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{\epsilon_{(i),k}^2}{2\sigma_i^2} \right) \right) & \text{if } \sigma_i \leq t\epsilon_{max} \\ 0 & \text{else} \end{cases} \tag{13}$$

which will give models 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 (13) is equivalent to removing models with $\sigma_i > t\epsilon_{max}$ from the candidate set prior to selection. The recommended way to enforce the constraint is to discard these candidates before setting up the matrix `Q` in order to speed up the optimization.

3.5 Prior Information

The Bayesian formulation of the problem allows to use additional information not present in the data, which enter the model in the form of the prior distribution. So far, we have assumed a uniform prior, i.e., the a-priori likelihood is the same for each explanation of the data. In some situations it may however be useful to impose a non-uniform prior. As already stated earlier, choosing a model is an interpretation of the data [15, 7]. There is no unique, "correct solution", the best choice depends on the model's purpose.

As an example, we will make use of a simple prior to increase the sensitivity to small motions. Geometric model selection is a powerful mechanism to prevent overfitting, i.e., it accepts a certain amount of residuals to make sure the model has the ability to predict unobserved data. If the task at hand is to separate small relative motions in the available data, without caring about unobserved correspondences, we may consider the same result an underfit with too large residuals – the reason not being that the observations have changed, but that our problem now is to find

small independent motions, not to find a compact description. The task to find small independent motions model selection must be biased so that

- the cost for an additional model is decreased compared to the benefit for the lower fitting error (so that a motion can more easily be split into two motions with lower error), and
- the cost for an additional model is decreased compared to the benefit for reducing outliers (so that a lower number of points justifies an additional model for previously unexplained matches).

The prior, which favors a description with a higher number m of independent motions, in this case works against the penalty terms of the criterion, which favor a description with few motions. Furthermore, the prior must be proportional to the total number of matches N_t, otherwise its influence will decrease $\rightarrow 0$ as the number of matches increases. The simplest prior with these properties is one with linear influence:

$$\mathcal{L}_{Pr}(m) = \frac{B^m}{\sum_{i=1}^{i=M} B^i} \quad \text{where} \quad B = C^{N_t} \tag{14}$$

In this expression, M is the maximum allowed number of motion models, but need not be known, because the denominator is a constant and can be dropped. The constant C determines the strength of the bias. Being part of the prior, it cannot be determined within the given Bayesian framework, but is an as yet arbitrary parameter, the choice of which requires external knowledge. Given that the model cost should be decreased by the prior, but remain > 0, the theoretical range is $1 < C < 4^2$. Writing $\lambda_4 = N_t \ln(C)$, the prior changes the diagonal elements of `Q` to

$$q_{ii} = N_i\lambda_1 - N_i \ln(\sigma_i^2) - E_i - \lambda_2 D_i - \lambda_3 K_i + \lambda_4 \tag{15}$$

The prior results in a constant benefit for each tentative model, which makes the model cheaper and, as desired, linearly decreases the total model cost with increasing m.

In section 4 we will give an example, how this prior works on a practical example. However, it shall be emphasized that priors should be used with caution – calling a shaky assumption a prior does not make it more respectable. Arbitrary priors, which are not based on dependable knowledge, are just the infamous "thresholds" or "damping factors" in Bayesian disguise.

4 Experiments

4.1 Synthetic Data

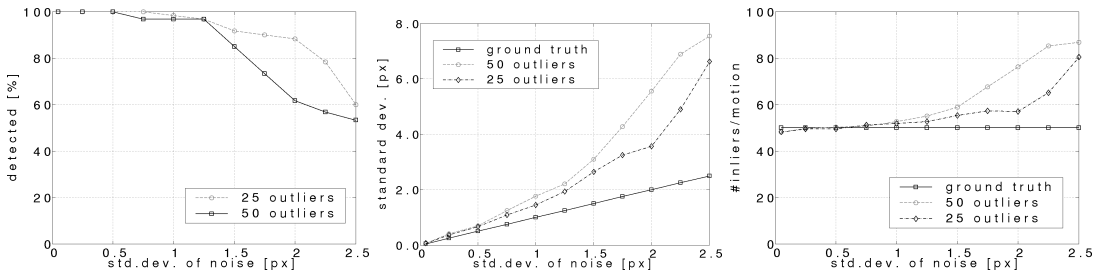
Experiments with random data were used to empirically assess the proposed method. The experiments assume a pair of images with 500×500 pixels. For the first experiment, spatially clustered clouds of 50 random points per model were generated on 1-3 randomly chosen motion models and perturbed with 0.5 pixel i.i.d. Gaussian noise, and 50 outliers were added from a uniform distribution over the two image planes. Then the algorithm was applied to the data, with 10000 initial candidate fundamental matrices and 2500 candidate homographies. The procedure was repeated 100 times. To judge the performance of the selection, the number and the type of recovered motions is used, while 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 number of points accepted as inliers and the estimate for the standard deviation grow, as more motions and more outliers are added, since outliers and overlap blur the borders between the distributions. In some cases one out of three motions was missed. This happened when two of the random models were very similar and have a large overlap. In this case, the cost for assigning the remaining points of the weaker model to he outliers is lower than the cost for an additional model. Whenever a motion was found, it was assigned the correct motion model.

The effect of merging similar motions is inevitable in the presence of outliers, since the outlier model allows for unexplained points. In other words, accepting outliers inherently reduces the ability to discriminate very similar models, because a certain number of inliers is needed to justify the cost of an additional model. The effect could be mitigated by a prior, which increases the cost of descriptions with many outliers – at the expense of spurious models in case of many real outliers. The result that the algebraic method of Vidal et al. [31] always finds the correct number of motion models can probably be attributed to the fact that the multibody fundamental matrix requires *every* point to satisfy one of the motion models. All detected motions were assigned the correct motion model.

number	**detected**	**correct**	**inliers**	σ **[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*

Table 1: 3D segmentation of random data. Left to right: number of true motion models present in the scene, percentage of detected motions, percentage of correctly selected motion models for the detected motions, average number of inliers per detected motion, average standard deviation of detected motions.

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 25 outliers. The amount of noise added to the inliers was increased from 0.05 to 2.5 pixels[1]. 30 tests were run at each noise level, again using 10000/2500 initial candidates. Since the ability to separate the two inlier distributions depends on the amount outliers, the whole test was also repeated with 50 outliers. The results are shown in Figure 3. Up to a noise level of 1.25 pixels (0.25% 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 ≈0.25% noise.



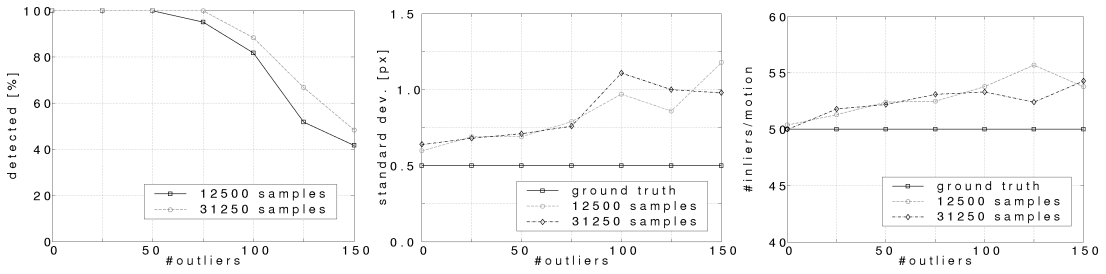
(a) (b) (c)

Figure 3: 3D segmentation at different noise levels. Left to right: mean percentage of detected motions, mean estimated standard deviation of motion models, mean number of inliers per model. See text for details.

In the third experiment, the number of motion models was again fixed to two and the measurement noise kept at 0.5 pixels, but the number of outliers was gradually increased. As expected, the method breaks down at a certain outlier percentage, which is due to the Monte-Carlo strategy. As the inlier fraction decreases, more and more samples are needed to obtain any correct candidate models for the selection process. When 75 real outliers (≈40%) are reached, which do not belong

[1]The minimal noise of 0.05 is required for the mean-shift procedure.

to any motion model, the method gradually break down. It can be seen from the estimated standard deviations and inlier numbers that more noise does not impair the 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 sample. The experiment was also repeated with a higher sample number of 25000/6250. As was to be expected, the results are slightly better, but on the whole the experiment confirms that the method can cope well with up to 40% of the entire data being outliers, before the results deteriorate. The results are summarized in Figure 4.



(a) (b) (c)

Figure 4: 3D segmentation with different amount of outliers. Left to right: mean percentage of detected motions, mean estimated standard deviation of motion models, mean number of inliers per model. See text for details.

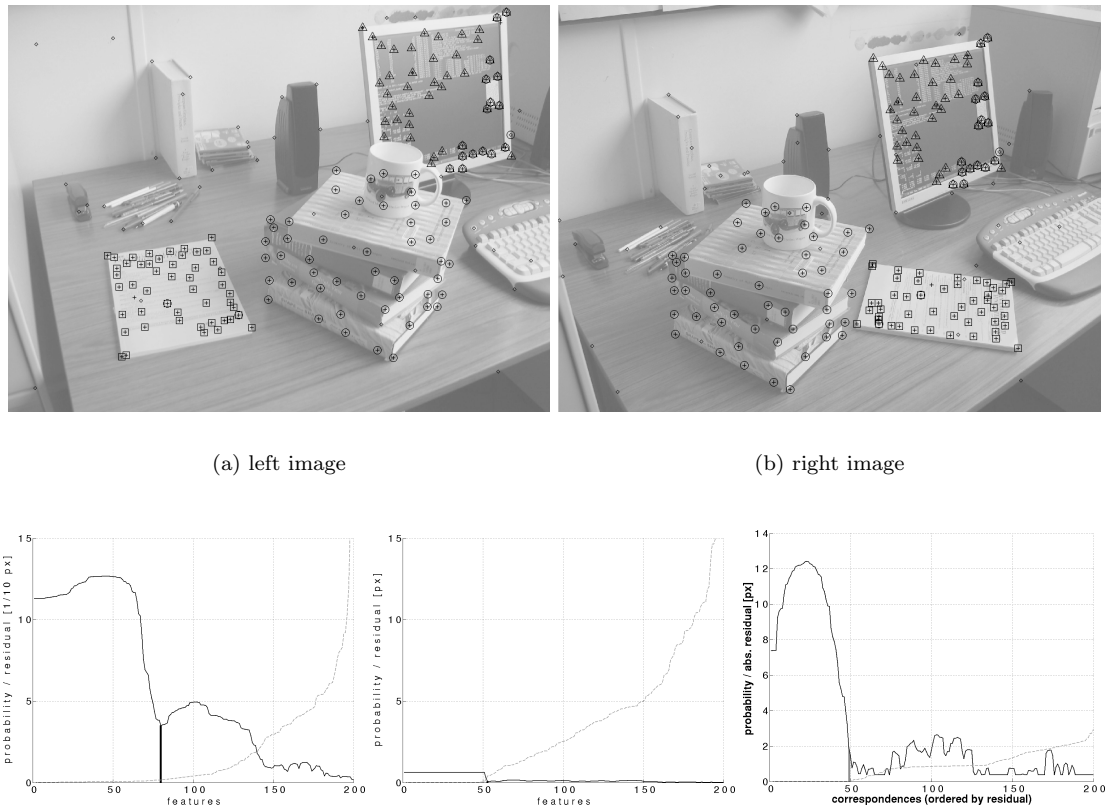
4.2 Real Data

For the first example, an image pair was taken with 3 different motions (1 fundamental matrix, 2 homographies). On each of the 3 regions, 50 correspondences were measured manually (because no wide-baseline matching routine was available). 50 spurious matches were added at apparent intersections, repetitive structures etc. The method is able to correctly detect the three motion models. 6400 fundamental matrices and 1600 homographies were initially sampled with the localized sampling scheme described in the previous section. Models with a standard deviation of $\sigma > 5$ pixels (because of the manual measurement and radial distortion) were discarded. The remaining candidates, 89 fundamental matrices and 34 homographies, were optimally fitted to the inliers and passed on to the model selection stage, which correctly retained 1 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. 98% of all inliers were assigned to the correct model.

The method allows for models to overlap, i.e., a match can be an inlier for two different motion models. We have not disambiguated these points. A common strategy is to assign each point to the model 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. The amount of overlap is given in the table. The overlap mainly consists of points which are not on the pile of books, but still satisfy the associated fundamental matrix, because of the less restrictive model. Geometrically, these points lie on corresponding epipolar lines of the motion, although they are not generated by it.

4.3 Using a prior

To demonstrate the influence of the prior given in section 3.5, we have applied our method to the last first and last image of the “car-truck-box” sequence also used by Vidal et al. [31, 30].



(a) left image (b) right image

(c) pdf for books motion (d) pdf for journal motion (e) pdf for screen motion

Figure 5: 3D segmentation of "desk" image pair. (a),(b): ground truth and segmentation result are overlayed on the images. small symbols denote the matches on different motions, large symbols the obtained clustering. Small diamonds are outliers. (c)-(e): absolute residuals (gray, dashed), probability density functions of absolute residuals (black, continuous), and estimated separation between inliers and outliers for the three selected models.

object	**motion**	**true**	**inliers**	**true inliers**	**other inliers**
books	F	50	69	50	19
journal	F	50	49	49	0
screen	H	50	49	48	1
outliers	—	50	*49*	*47*	*2*

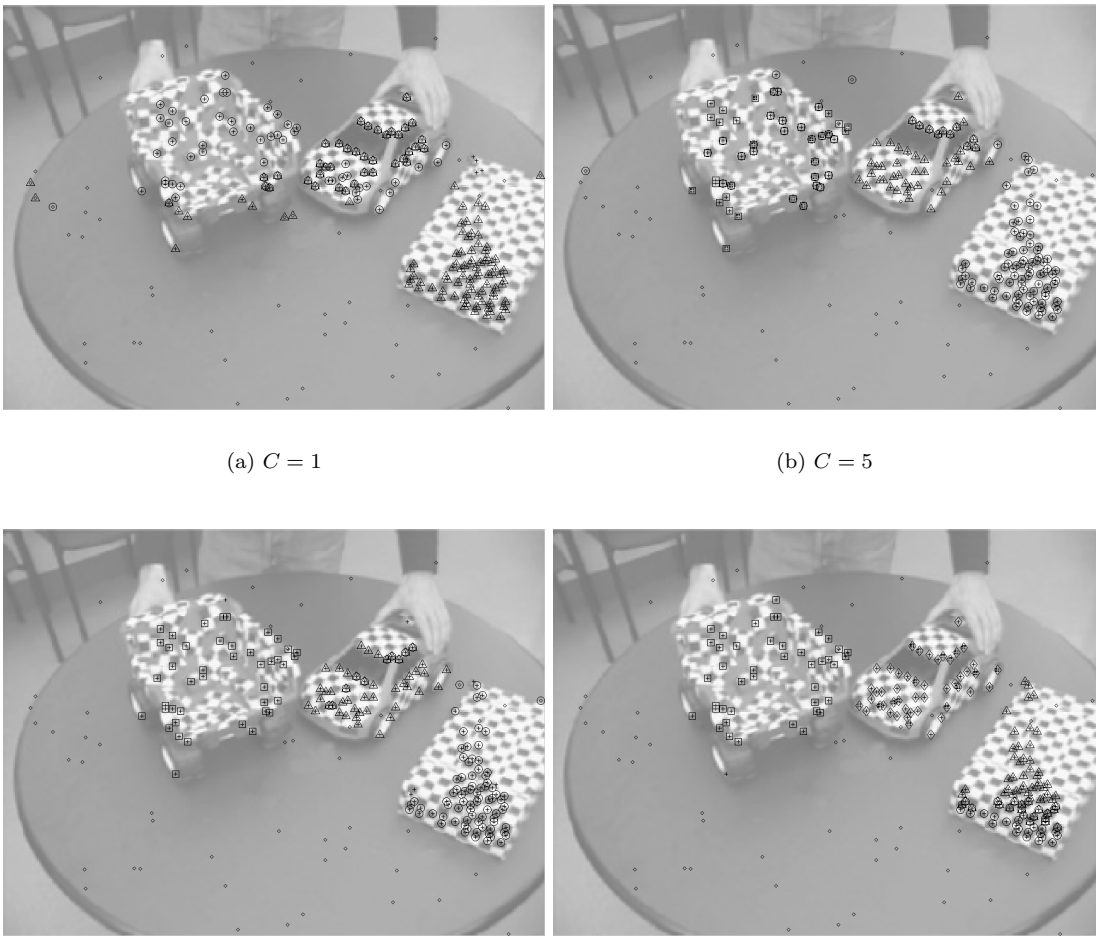
Table 2: 3D segmentation results for "desk" image pair. The outliers are a rejection class for points not assigned to any model.

The dataset contains 3 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 non-planar objects with general motion. However the average Sampson residual when fitting a fundamental matrix to the matches on the box is $s_{F,b} = 0.53$ pixels, while the Sampson residual when fitting a fundamental matrix to *the car and the truck together* is only $s_{F,ct} = 0.15$ pixels. Furthermore, the two motions are so small that the average Sampson error for fitting a homography is $s_{H,c} = 0.13$ pixels for the car and $s_{H,t} = 0.44$ pixels for the truck, as compared to $s_{F,c} = 0.07$ and $s_{F,t} = 0.11$ for a fundamental matrix.

To test our method, 50 outliers where added by sampling spurious matches from a uniform

distribution. Then the method was applied to the data with uniform prior, and with the prior from equation (15), using the values $C = 5, 10, 15, 16$. The results are depicted in Figure 6. Without prior information, 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. If the non-uniform prior is introduced with $C = 5$, the cost for an additional model decreases. The motions of the car and the truck are separated and assigned two distinct homographies. The same result is obtained with $C = 10$, while with $C = 15$ the cost for an additional model has decreased some much that the two planes of the box are also assigned two homographies, which is a different valid interpretation of the same data. The theoretical maximum $C = 16$ reduces the model cost so far that even a single non-overlapping point sometimes justifies a new model, leading to extreme overfitting.

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. The results are shown in Figure 6.



(a) $C = 1$ (b) $C = 5$

(c) $C = 10$ (d) $C = 15$

Figure 6: 3D segmentation of "cars" image pair. Left image and segmentation results with different priors. Small symbols denote matches on different motions, large symbols the obtained clustering. Small diamonds are outliers. The results are given for a uniform prior (a), and biased towards higher sensitivity with $C = 5$ (b), $C = 10$ (c), and $C = 15$ (d). Details are given in the text.

5 Concluding Remarks

We have presented a scheme for robust multibody structure-and-motion in the presence of different motion models, noise of unknown standard deviation, and outliers. The method follows the recover-and-select paradigm: candidate motion models are instantiated through Monte-Carlo sampling, and for each candidate the inlier set and the corresponding standard deviation are estimated by applying mean-shift to recover a non-parametric probability density function of the residuals. With this information, the redundant set of candidates is pruned through geometric model selection by maximizing the posterior likelihood of the description.

The method needs no thresholds to establish the tentative models and to select the optimal set, except for an upper bound of the allowable measurement error of an inlier. However the random sampling does rely on a heuristic localization scheme to keep the number of required samples in a manageable order of magnitude. The method has been successfully applied to different data sets.

The basic ideas behind the method are generic for fitting multiple models 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 large number of unknowns, and the need to fit manifolds of different dimension. The application to problems such as curve fitting or range segmentation is straight-forward.

Acknowledgments

I would like to thank Hanzi Wang for helping me to get started with the TSSE-estimator, Horst Bischof and Ales Leonardis for the code to solve the quadratic boolean optimization problem, and Rene Vidal for providing the “car-truck-box” data.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Proc. 2nd International Symposium of Information Theory*, pages 267–281, 1973.

[2] S. Avidan and A. Shashua. Trajectory triangulation: 3d reconstruction of moving points from a monocular image sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(22):348–357, 2000.

[3] G. L. Bretthorst. An introduction to model selection using probability theory as logic. In G. R. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*, pages 1–42. Kluwer Academic Publishers, 1996.

[4] D. Comaniciu and P. Meer. Mean shift: A robust approach towards feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 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? In *Proc. 2nd European Conference on Computer Vision, Sta. Margherita Ligure, Italy*, pages 563–578, 1992.

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

[8] F. Glover and M. Laguna. Tabu search. In C. Reeves, editor, *Modern Heuristic Techniques for Combinatorial Problems*. Blackwell Publishers, 1993.

[9] M. Han and T. Kanade. Reconstruction of scenes with multiple linearly moving objects. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, pages 542–549, 2000.

[10] R. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *Proc. 2nd European Conference on Computer Vision, Sta. Margherita Ligure, Italy*, pages 579–587, 1992.

[11] R. Hartley and R. Vidal. The multibody trifocal tensor: Motion segmentation from 3 perspective views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Washington, D.C.*, volume 1, pages 769–775, 2004.

[12] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[13] P. J. Huber. *Robust Statistics*. John Wiley and Sons, 1981.

[14] K. Kanatani. *Statistical Optimation for Geometric Computation : Theorie and Practice*. North Holland Elsevier, 1996.

[15] K. Kanatani. Geometric information criterion for model selection. *International Journal of Computer Vision*, 26(3):171–189, 1998.

[16] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric parametric models. *International Journal of Computer Vision*, 14(1):253–277, 1995.

[17] C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[18] Yi Ma, J. Kosecka, S. Soatto, and S. Sastry. *An invitation to 3-D vision*. Springer Verlag, 2003.

[19] S. J. Maybank and P. F. Sturm. MDL, collineations and the fundamental matrix. In *Proc. 10th British Machine Vision Conference, Nottingham, UK*, 1999.

[20] M. Pollefeys, F. Verbiest, and L. Van Gool. Surviving dominant planes in uncalibrated structure and motion recovery. In *Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark*, 2002.

[21] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory*, pages 629–636, 1984.

[22] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons, 1987.

[23] G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:497–511, 1978.

[24] A. Shashua and A. Levin. Multi-frame infinitesimal motion model for the reconstruction of (dynamic) scenes with multiple linearly moving objects. In *Proc. 8th International Conference on Computer Vision, Vancouver, Canada*, pages 592–599, 2001.

[25] M. Stricker and A. Leonardis. ExSel++: A general framework to extract parametric models. In *Proc. Computer Analysis of Images and Patterns*, pages 90–97, 1995.

[26] P. Sturm. Structure and motion of dynamic scenes – the case of points moving in planes. In *Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 867–882, 2002.

[27] P. H. S. Torr. Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society of London A*, 356(1740):1321–1340, 1998.

[28] P. H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1):35–61, 2002.

[29] B. Triggs. A new approach to geometric fitting. Available from `http://www.inrialpes.fr/movi/people/Triggs`, 1998.

[30] R. Vidal and S. Sastry. Optimal segmentation of dynamic scenes from two perspective views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin*, 2003.

[31] R. Vidal, S. Soatto, Yi Ma, and S. Sastry. Segmentation of dynamic scenes from the multibody fundamental matrix. In *Proc. ECCV Workshop on Visual Modeling of Dynamic Scenes*, 2002.

[32] H. Wang and D. Suter. MDPE: A very robust estimator for model fitting and range image segmentation. *International Journal of Computer Vision*, 59(2):139–166, 2004.

[33] H. Wang and D. Suter. Robust fitting by adaptive-scale residual consensus. In *Proc. 8th European Conference on Computer Vision, Prague, Czech Republic*, pages 107–118, 2004.

[34] L. Wolf and A. Shashua. Two-body segmentation from two perspective views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, pages 263–270, 2001.

[35] Z. Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59–76, 1997.