Two-view Multibody Structure-and-Motion with Outliers

Konrad Schindler and David Suter
Electrical and Computer Systems Engineering
Monash University, Clayton, 3800 Victoria, Australia

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 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 to find an optimal set of motion models for the measurements is solved through Monte-Carlo sampling, careful statistical analysis of the data and simultaneous selection of multiple motion models.

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 [7, 4]. 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 [4]. 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* [1]. To decide between different motion models, 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 [9], who also first recognized that the dimension of the fitted manifold requires separate treatment [8].

[1] 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 general and can be extended to other, simpler models, as for example shown in [16].

Soon after the main SaM-theory had been established, researchers turned to the more challenging case of *dynamic* scenes, e.g. [22]. Recently an excellent extension of algebraic SaM-theory to dynamic scenes has been presented [19]. 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 method has been expanded to a "multibody homography" [18], but it is not designed to mix different motion models, and it does not include an outlier model. The latter, together with the non-linear nature of the problem, makes the purely algebraic approach 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. 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. Torr has proposed an iterative strategy [16]: 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. Therefore, the result has to be post-processed with expectation maximization and model pruning. In iterative MSaM, the 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, non-parametric statistical analysis of the residuals is used to individually estimate the scale of the noise for each model. Given the scale and the number of inliers found with this scale, the likelihood of each model can be computed, which is then used as a measure of the goodness-of-fit during model selection.

There are two original contributions in this paper, one in each step. Firstly, 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 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 wider than the distribution, then the number of inliers and the fitting residuals are over-estimated; if the threshold is too narrow, the two quantities are under-estimated. The incorrect estimates will influence model selection, because these quantities are exactly the variables used to assess the goodness-of-fit. In contrast, the new method estimates the residual distribution for each tentative model and computes an individual standard deviation from it.

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

2 Generating candidate models

Sampling. For 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, 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 2 motions are present, and becomes intractable for more than 2 motions. A practical 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 one. 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 a reasonable compromise between local coherence and global extension, which works well for different images. 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. 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%. Hence, if an independently moving object covers at least 10% of the image, and is not very elongated in shape, there will be at least one region, in which the object covers ≈50% of the entire area. Details can be found in [13].

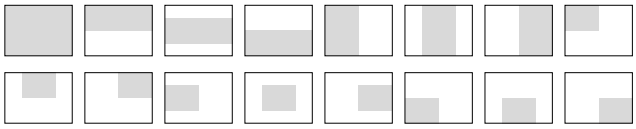

Figure 1. Local sampling scheme for tentative motion models. Samples are drawn from sub-regions of the image plane to exploit spatial coherence and reduce the required sample number.

Estimating standard deviations. Given a model and the Sampson-residuals [4] of the data points, the scale of the noise can be estimated without any further knowledge by applying the MDPE-estimator of Wang and Suter [20]. In a nutshell, the probability density function (*pdf*) of the ordered absolute residuals is estimated with a kernel density method, and the extrema of the *pdf* can be directly found with the mean-shift algorithm [3], without explicitly having to recover the entire function. The bandwidth for the mean-shift algorithm can be selected automatically from the data [21]. Due to lack of space, we refer the interested reader to the original publication for details.

Assuming that the inliers have mean zero, the valley of the *pdf* closest to zero is a sensible point to separate inliers from outliers. Points with a residual lower than the valley are retained as inliers and their standard deviation is computed. The procedure is illustrated in Figure 3 for one of the motions.

Estimating the variance and inlier threshold of each model separately from the data considerably improves the power of the method, compared with a fixed threshold between inliers and outliers. When searching for *multiple* models rather than a single one, an incorrect threshold can severely impair the results: if the threshold is too low, only a subset of the inliers is found and assigned to the model, and the remaining inliers may give rise to a second, similar model, leading to overfitting. If, on the contrary, the threshold is too high, the overlap between models will be

overestimated. This can lead to underfitting, because one of the models will claim too many of the data points, leaving only little support for the second model. Most important, wrong estimates for the standard deviation and the number of inliers lead to a wrong estimate of the model's likelihood and adversely affect model selection.

As noted by Rousseeuw [12] and confirmed by other authors, 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 inliers.

3 Model Selection

Principle of Geometric Model Selection. To select the optimal set of models, 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 depending on their dimension and cardinality. There are several 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 [14], an approximation of the a-posteriori likelihood, and Rissanens *minimum description length* MDL [11], an information theoretic criterion that seeks to minimize the coding length of the data. In spite of their 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 – 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 [8]. GIC selects the model $\mathcal{M}$ which maximizes

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

where N_t is the total number of correspondences, 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, 3 for a fundamental matrix). $\mathcal{L}$ is the likelihood of the model. 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 [16, 17]. GRIC selects the model $\mathcal{M}$ which maximizes

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

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 [9, 5]. We agree with this view, and in fact will show that one can construct a prior which converts one criterion int othe other. 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, changing only the penalty terms.

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 [2], 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 all models established previously, the total likelihood has to be split into the contributions from the single models. At the same time, we have to account for the fact that models may overlap, i.e., data points may be inliers to more than one model. Points in the overlap should contribute only once to the overall likelihood. 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 2 models, but the number of these points is small compared to those satisfying exactly 2 models. The approximation is necessary to yield a tractable optimization problem, as explained later in this section.

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

$$\mathcal{L}_i = \prod_{k\in\mathcal{V}_i} \left(\frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{\epsilon^2_{(i),k}}{2\sigma_i^2} \right) \right) = \prod_{k\in\mathcal{V}_i} \mathtt{G}^{(i)}_k \tag{3}$$

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 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

$$\mathcal{L}_{[ij]} = \prod_{k\in\mathcal{V}_{[ij]}} \min(\mathtt{G}_k^{(i)}, \mathtt{G}_k^{(j)}) = \prod_{k\in\mathcal{V}_{[i]}} \mathtt{G}_k^{(i)} \prod_{k\in\mathcal{V}_{[j]}} \mathtt{G}_k^{(j)} \quad (4)$$

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 outliers w.r.t. $\widehat{\mathcal{C}}$ by $\mathcal{L}_{/\widehat{\mathcal{C}}}$, then the total likelihood of $\widehat{\mathcal{C}}$ is

$$\mathcal{L}_{\widehat{\mathcal{C}}} = \mathcal{L}_{/\widehat{\mathcal{C}}} \prod_{i\in\widehat{\mathcal{C}}} \mathcal{L}_i \prod_{i,j\in\widehat{\mathcal{C}}} \mathcal{L}_{[ij]}^{-1} \quad (5)$$

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, and the likelihood of h outliers is $\mathcal{L}_{/\widehat{\mathcal{C}}} = P^h$. If the search area for matching is restricted, A has to be changed appropriately.

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. Then the log-likelihood of the chosen subset is

$$\ln(\mathcal{L}) = \sum_{i\in\mathcal{C}} (b_i \ln(\mathcal{L}_i)) - \sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{C}} \left(b_i b_j \ln\left(\mathcal{L}_{[ij]}\right)\right) + h\ln(P) \quad (6)$$

In this expression we can substitute the likelihoods with expressions (3) and (4), 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 drop the constant terms which will not unfluence maximization. After some manipulations (details are given in [13]) this leads to

$$\begin{aligned} 2\ln(\mathcal{L}) = &\sum_{i\in\mathcal{C}} (b_i(N_i\lambda_1 - N_i\ln(\sigma_i^2) - E_i)) - \\ &\sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{C}} (b_i b_j (N_{[ij]}\lambda_1 - N_{[i]}\ln(\sigma_i^2) - \\ &E_{[i]} - N_{[j]}\ln(\sigma_j^2) - E_{[j]})) \end{aligned} \quad (7)$$

where $\lambda_1 = -2\ln(P) - \ln(2\pi)$ and the sums of squared errors $E_i = \frac{1}{\sigma_i^2} \sum_{k\in\mathcal{V}_i} \epsilon_{(i),k}^2$ and $E_{[i]} = \frac{1}{\sigma_i^2} \sum_{k\in\mathcal{V}_{[i]}} \epsilon_{(i),k}^2$.

Maximizing the Criterion. Previously, model selection criteria have either been used to select an unknown number of models with the same dimension at once, such as in [10], or to select one model of varying dimension at a time, as in [8, 16]. 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.

With expression (7) for the likelihood, the GRIC (2) for a model collection $\widehat{\mathcal{C}}(\mathbf{b})$ can be written as

$$\mathrm{GRIC}(\mathbf{b}) = \mathbf{b}^\mathsf{T}\mathtt{Q}\mathbf{b} \quad (8)$$

where $\mathtt{Q}$ is a symmetric matrix [10]. Let the constants $\lambda_2 = N_t\ln(4)$ and $\lambda_3 = \ln(4N_t)$. Then the diagonal elements of $\mathtt{Q}$ are

$$q_{ii} = N_i\lambda_1 - N_i\ln(\sigma_i^2) - E_i - \lambda_2 D_i - \lambda_3 K_i \quad (9)$$

and the off-diagonal elements, which handle the overlap between different tentative models, are

$$\begin{aligned} &q_{ij} = q_{ji} = \\ &-\frac{1}{2}\left(N_{[ij]}\lambda_1 - N_{[i]}\ln(\sigma_i^2) - E_{[i]} - N_{[j]}\ln(\sigma_j^2) - E_{[j]}\right) \end{aligned} \quad (10)$$

Intuitively, equation (9) favors motions which reduce the number of outliers (large N_i), have low error (low σ_i and E_i), and have low dimension D_i and parameter count K_i. Note that no parameters have to be tuned in (9) and (10).

Maximizing expression (8) over $\mathbf{b}$ is an NP-hard combinatorial problem. Taboo-search [6] is a standard method for approximate solution of such problems. Roughly speaking, Taboo-search performs gradient ascent through switching on or off elements of $\mathbf{b}$, but does *not* stop at the first local minimum. Instead, the search continues such that recent moves are not reversed, ensuring that it departs far enough from a local minimum. For details see for example [15].

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, 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 advisable to

use a more conservative upper bound $t\epsilon_{max}, t \approx 2$. To formally add it to the probabilistic formulation of the optimization problem, one would have to redefine the likelihood (3) of a candidate model $\mathcal{V}_i$ as

$$\mathcal{L}_i = \begin{cases} \prod_{k \in \mathcal{V}_i} \left(\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{\epsilon^2_{(i),k}}{2\sigma_i^2} \right) \right) & \text{if } \sigma_i \leq t\epsilon_{max} \\ 0 & \text{else} \end{cases} \tag{11}$$

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 (11) is equivalent to removing models with $\sigma_i > t\epsilon_{max}$ from the candidate set prior to selection. The latter speeds up the optimization.

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, both GRIC and GIC sometimes do not give satisfactory results if the task is to segment small relative motions. The reason is a different 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 model, so we are in fact *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 a Bayesian framework, information not manifest in the data is introduced in the form of the prior distribution. The penalty terms in the criterion are a prior, which expresses the belief that a simpler description of the data is more likely. The new prior term shall mitigate this, saying that it is "not that much more likely". It must be proportional to the total number of matches N_t, otherwise its influence will decrease $\rightarrow 0$ as the number of matches increases[2]. The simplest prior with these properties, which preserves the ratio between the penalties for a fundamental matrix and a homography, is

$$\mathcal{L}_{Pr} = \frac{1}{S_{Pr}} \prod_{i \in \widehat{\mathcal{C}}} C^{B_i N_t} \ , \ B_i = \begin{cases} \mathtt{H} : 1 \\ \mathtt{F} : \frac{3\ln(4)N_t + 7\ln(4N_t)}{2\ln(4)N_t + 8\ln(4N_t)} \end{cases} \tag{12}$$

S_{Pr} is the combinatorial sum over all possible $\mathcal{L}_{Pr}$, but need not be known, because it is 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 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 $(1 < C < 4^2)$. Writing $\lambda_4 = N_t \ln(C)$, the prior changes the diagonal elements of $\mathtt{Q}$ to

$$q_{ii} = N_i\lambda_1 - N_i \ln(\sigma_i^2) - E_i - \lambda_2 D_i - \lambda_3 K_i + B_i\lambda_4 \tag{13}$$

As desired, the penalties for new models have been decreased, treating all motions in an equal way independent of the total number, and preserving the ratio between model penalties. In section 4 the effect of this prior is shown on a practical example.

The prior likelihood (12) is only the simplest representative of a more general prior

$$\mathcal{L}_{Pr} = \frac{1}{S_{Pr}} \prod_{i \in \widehat{\mathcal{C}}} C^{f(N_t)} \tag{14}$$

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

$$f(N_t) = \begin{cases} \mathtt{H} : 2(\ln(4) - 2)N_t + 8\ln(4N_t) - 16 \\ \mathtt{F} : 3(\ln(4) - 2)N_t + 7\ln(4N_t) - 14 \end{cases} \tag{15}$$

results in a prior, which converts GRIC into GIC. However, it remains to be investigated, how the function f could be selected in a useful and 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 Experiments

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 estimates for the models' standard deviations grow, as more motions are added, since pseudo-outliers on other motions and overlap blur the borders between the distributions. In some cases one out of three motions was missed. This happens when

[2]GRIC/GIC 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.

two of the random models are very similar and have a large overlap, so that the cost for assigning the remaining points of the weaker model to the outliers is lower than the cost for an additional model. This effect is inevitable in the presence of outliers: allowing for unexplained points inherently reduces the ability to discriminate similar models. 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.

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 motion models, detected motions, correctly classified motions, average number of inliers, average standard deviation.

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[3]. 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 25 outliers. The results are shown in Figure 2. 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.

The third experiment again used 2 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 (≈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 sample. The experiment was also repeated with a higher sample number of 25000/6250. The results are slightly better, but on the whole they confirm that the method can cope with up to ≈40% outliers. The results are summarized in Figure 2.

We have also tested the proposed method on a real image pair with 3 different motions. On each of the 3 regions, 50 correspondences were measured manually. 50 spurious matches were added at apparent intersections, repetitive structures etc. 6400 fundamental matrices and 1600 homographies were initially sampled with the sampling scheme described in the previous section. Of these candidates, 89 fundamental matrices and 34 homographies survived the constraint ($\sigma_i < 4$ pixels) and were 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.

We have not disambiguated points which satisfy more than one model. 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 overlap mainly consists of points which are not on the pile of books, but still satisfy the associated fundamental matrix, because it is less restrictive than a homography.

object	motion	true	inliers	corr. inliers
books	F	50	69	50
journal	H	50	49	49
screen	H	50	49	48
outliers	—	50	*49*	*47*

Table 2. 3D segmentation results for "desk" images. The outliers are a rejection class for points not assigned to any model. See text for details.

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. [19, 18]. 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 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 aver-

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

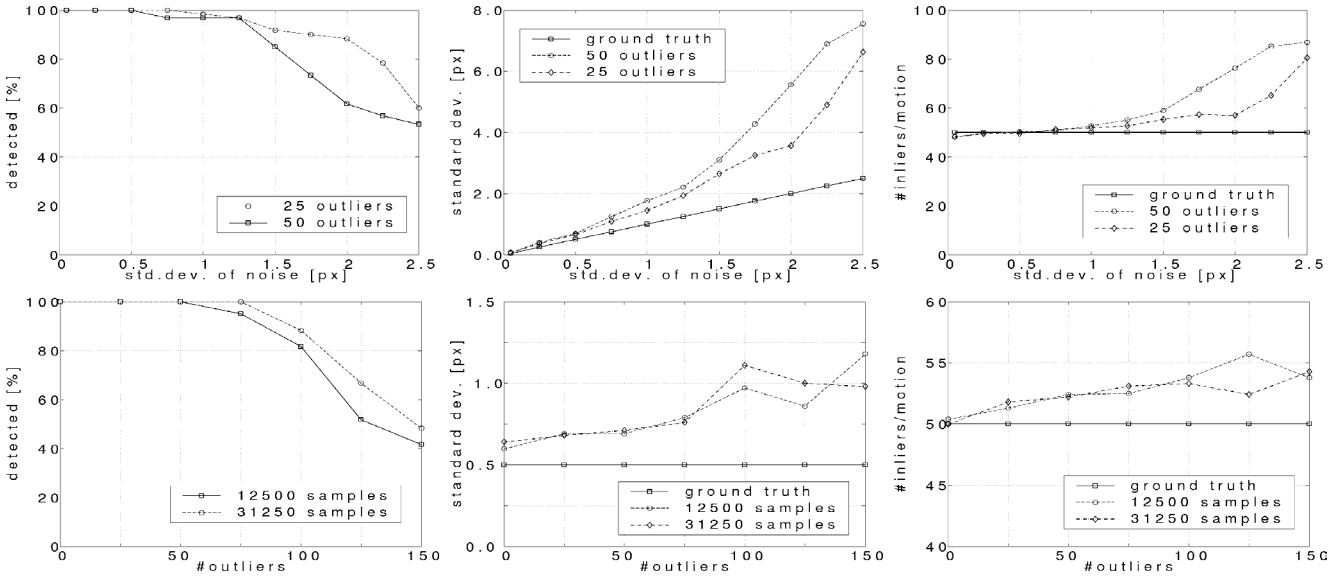

Figure 2. 3D segmentation with synthetic data. Top row: results at different noise levels. Bottom row: results with different amount of outliers. See text for details.

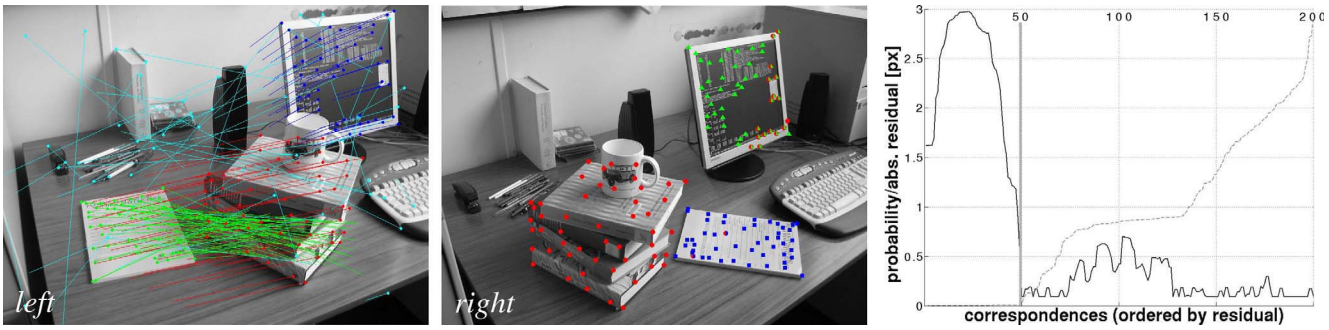

Figure 3. 3D segmentation results for "desk" images. Left: true motion overlayed in left image. Center: segmentation overlayed on right image. Circles denote fundamental matrices, polygons are homographies. Right: absolute residuals (gray, dashed), probability density function (black, continuous), and separation between inliers and outliers for the screen.

age 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.

50 outliers were added by sampling spurious matches from a uniform distribution. Then the method was applied to the data, using the prior from equation (13) with different values for C. The results are depicted in Figure 4. With a uniform prior $C = 1$, 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 $C = [5 \dots 6]$, the motions of the car and the truck are separated and assigned two homographies. With $C = [7 \dots 12]$, the truck is assigned a fundamental matrix instead, and with $C = 13$ each motion is modelled 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.

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.

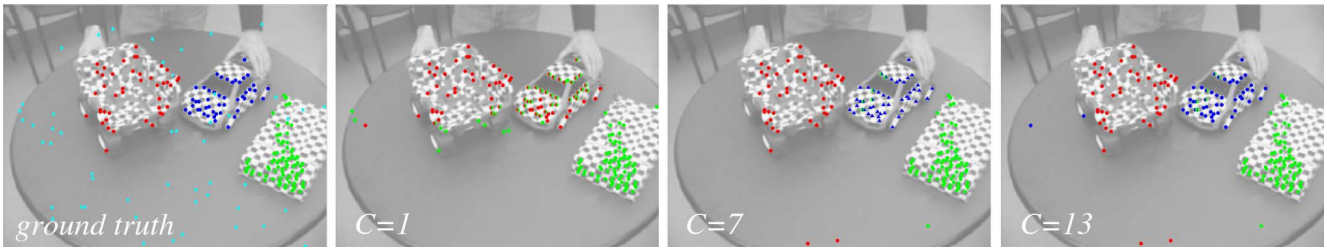

Figure 4. 3D segmentation of "cars" image pair using priors of different strength. Left to right: ground truth, uniform prior ($C = 1$), weak prior ($C = 7$), strong prior ($C = 13$). Circles denote points on a fundamental matrix, triangles are points on a homography, diamonds are outliers. Details are given in the text.

The method simultaneously recovers all present motions and needs no thresholds, except for an upper bound of the allowable measurement error. However random sampling does rely on a heuristic local scheme to keep the number of required samples in a manageable order of magnitude.

The underlying ideas are generic for robustly 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 need to fit up to three-dimensional manifolds, and to decide between manifolds of varying dimension.

The use of non-uniform priors has been briefly discussed to adapt the method to different vision tasks, but needs to be investigated in more detail.

Acknowledgments

We would like to thank Hanzi Wang, Horst Bischof, and Ales Leonardis for source code, and Rene Vidal for the "car-truck-box" data. This work was carried out within the *Institute for Vision Systems Engineering*.

References

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

[2] 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.

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach towards feature space analysis. *IEEE TPAMI*, 24(5):603–619, 2002.

[4] O. Faugeras, Q.-T. Luong, and T. Papadopoulo. *The geometry of multiple images*. MIT Press, 2001.

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

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

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

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

[9] K. Kanatani. Geometric information criterion for model selection. *IJCV*, 26(3):171–189, 1998.

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

[11] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Trans. Information Theory*, 30:629–636, 1984.

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

[13] K. Schindler. Simultaneous, robust fitting of multiple 3D motion models. Technical report MECSE-12-2004, Monash University, 2004.

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

[15] 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.

[16] P. H. S. Torr. Geometric motion segmentation and model selection. *Phil. Trans. Royal Society A*, 356(1740):1321–1340, 1998.

[17] P. H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *IJCV*, 50(1):35–61, 2002.

[18] R. Vidal and Yi Ma. A Unified Algebraic Approach to 2-D and 3-D Motion Segmentation. In *Proc. 8th ECCV*, pages 1–15, 2004.

[19] 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.

[20] H. Wang and D. Suter. MDPE: A very robust estimator for model fitting and range image segmentation. *IJCV*, 59(2):139–166, 2004.

[21] H. Wang and D. Suter. Robust fitting by adaptive-scale residual consensus. In *Proc. 8th ECCV*, pages 107–118, 2004.

[22] L. Wolf and A. Shashua. Two-body segmentation from two perspective views. In *Proc. IEEE CVPR*, pages 263–270, 2001.