**Noname manuscript No.**
(will be inserted by the editor)

# Efficient and Robust Model Fitting with Unknown Noise Scale

**Stuart B. Heinrich**

the date of receipt and acceptance should be inserted later

**Abstract** This paper addresses the general problem of robust parametric model estimation from data that has both an unknown (and possibly majority) fraction of outliers as well as an unknown scale of measurement noise. We focus on computer vision applications from image correspondences, such as camera resectioning, estimation of the fundamental matrix or relative pose for 3D reconstruction, and estimation of 2D homographies for image registration and motion segmentation, although there are many other applications. In practice, these methods typically rely on predefined inlier thresholds because automatic scale detection is usually too unreliable or too slow. We propose a new method for robust estimation with automatic scale detection that is faster, more precise and more robust than previous alternatives, and show that it can be practically applied to these problems.

**Keywords** robust estimation · RANSAC · scale estimation · structure from motion

## 1 Introduction

Data fitting (i.e., estimating the parameters of some hypothetical model that best explains a set of data measurements) is a critical task that arises in many disciplines. In general, the measurement set will contain some unknown fraction of outliers, and the good measurements will be subject to some unknown scale of noise, typically assumed to be Gaussian.

Least squares (LS) estimators, which minimize the sum of squared residual errors, are efficient and optimal for Gaussian noise (Huber, 1972), but are highly sensitive to outliers (Wang and Suter, 2004a). The *breakdown point* of the LS estimator is 0% because the estimate may be arbitrarily skewed when the percentage of outliers is greater than 0% (Rousseeuw, 1987, p.9). A more robust approach is the least median of squares (LMS) estimator (Rousseeuw, 1984), which minimizes the median of squared residuals, and has a breakdown point of 50%.

Accumulator based methods (e.g., the Hough transform (Hough, 1962; Duda and Hart, 1972) or variations such as the Randomized Hough Transform (RHT) (Xu et al., 1990)) have no specific breakdown point, and are popular for simple line and curve detection. However, they are non-optimal, and are limited in their general applicability because they require discretization of a $p$-dimensional space for models with $p$ parameters, which would result in prohibitively high time and space complexity for many problems.

Perhaps the most well-known and generally applicable robust estimator without a breakdown point is RANdom SAmple Consensus (RANSAC) (Fischler and Bolles, 1981). RANSAC uses a hypothesize-and-test framework by randomly sampling subsets of the measurement set, and retaining the model that maximizes the number of inliers according to some threshold. Because it does not require discretization of the search space, estimation of high-dimensional models is computationally feasible, and there is no breakdown point beyond the minimal fraction necessary to define a model.

Due to its success, there have been many RANSAC variations. To summarize, robust M-estimators (Huber, 1981) were used for model evaluation with MSAC and MLESAC (Torr and Zisserman, 2000), the inner optimization from LO-RANSAC (Chum et al., 2003) attempts to compensate for the unlreaistic assumption that all models estimated from uncontaminated (albeit noisy) data are good, and explicit testing for degenerate configurations has been incorporated in DEGENSAC (Chum et al., 2005) and QDEGSAC (Frahm and Pollefeys, 2006). Other improvements have focused on performance optimizations by using heuristic bail-out tests (Matas and Chum, 2004; Capel, 2005; Chum and Matas, 2008) or guided sampling as in PROSAC (Chum and Matas, 2005), Preemptive RANSAC (Nister, 2003), and ARSSAC (Raguram et al., 2008). See Raguram et al. (2008) for a more thorough survey.

All of these aforementioned RANSAC variations implicitly require accurate *a priori* knowledge of the scale of inlier noise in order to choose the threshold. In many cases, it is possible to choose a reasonable threshold based on domain knowledge, but the sensitivity to this choice is undesirable and can sometimes result in instability. Indeed, it has been often noted that given a bad choice of threshold, RANSAC will completely break down (Wang and Suter, 2004a,b).

S. Heinrich
North Carolina State University, Box 8206, Raleigh, NC 27695
E-mail: sbheinri@ncsu.edu

One of the first approaches attempting to overcome this limitation was to first make a robust estimate of the model using LMS and then make a robust estimate of scale using the median squared residual (Rousseeuw, 1987), as proposed in Torr and Murray (1997). However, because LMS and the median scale estimate both have 50% breakdown points, this method cannot be applied to data sets with more than 50% outliers, as is often the case.

More recent RANSAC variations have attempted to incorporate scale estimation with model estimation. For example, ASSC (Wang and Suter, 2004a) modified the RANSAC objective by maximizing the inlier count divided by a robust estimate of scale. However, there is no statistical support for this modified objective, and it does not always detect the correct scale. ASSC also retains adaptive sampling from RANSAC, but not in a statistically valid way, and this can lead to premature convergence at grossly over-estimated scales when the outlier ratio is high. A similar approach was taken in Fan and Pylvänäinen (2008), who suggest modifying the objective function to seek the model which minimizes their proposed weighted median absolute deviation (WMAD) estimate of scale.

Projection-based M-estimators (pbM-estimators) were used in the 'projection pursuit' approach of Chen and Meer (2003), and some performance enhancements were proposed in Rozenfeld and Shimshoni (2005); Subbarao and Meer (2005, 2006). The most recent and best-performing technique along these lines is ASKC (Wang et al., 2010).

ASKC is an improvement upon the original ASSC algorithm, with a more statistically motivated objective function. The basic idea is to choose the random model hypothesis that maximizes a kernel density estimate in residual space centered at the origin. ASKC also abandons the earlier attempt at adaptive sampling from ASSC, and instead uses a fixed number of samples. The recognition that adaptive sampling does not work in this context is a significant limitation in comparison to the original RANSAC algorithm, because using a fixed number of samples either prevents good performance in the presence of high inlier ratios (excessive sampling), or induces a breakdown point in the presence of low inlier ratios due to not enough sampling to find the structure within the data.

A subtle theoretical problem with ASKC is that the method was derived based on the assumption that the residual distribution should be normal and hence have a mode at the origin, but in their experiments is often applied to the distribution of squared fitting errors which has, in general, a scaled $\chi^2_k$-distribution with a non-central mode that depends on $\sigma$ for $k > 2$ (Section 3).

Ultimately, the greatest limitation of both ASSC and ASKC is that they attempt to estimate the scale directly from the fully contaminated set of residuals, which is an inherently difficult problem to solve under high outlier ratios. The proposed Two-Step Scale Estimator (TSSE) from Wang and Suter (2004a), which is used by both methods, relies on a kernel density estimate (KDE) of the residual error distribution and is capable of functioning under high outlier ratios, but only if the kernel bandwidth is chosen properly. Automatic methods for choosing the bandwidth rely on an accurate estimate of scale. Thus, it is somewhat of a 'chicken-and-egg' problem.

Wang et al. (2010) propose obtaining the initial estimate using the $k$-th order statistic, which we find sometimes works and sometimes does not. Outlier contamination can lead to over-estimated bandwidth, in turn leading to oversmoothing of the KDE, and finally poor scale estimation with TSSE. To counteract this oversmoothing they propose using a fraction $c_h \in (0, 1)$ of the automatically derived bandwidth, but we find that there is no 'one size fits all' value of this parameter, because it depends largely on the scale and distribution of outliers. In summary, the overall sensitivity of ASKC to the scale estimator leaves us unsatisfied.

Another recent approach to automatic scale estimation is based on the recognition that RANSAC tends to exhibit the greatest consistency in the discovered models when the threshold is set near the true scale level. This observation was first exploited in StaRSaC (Choi and Medioni, 2009), which performs a brute force search across a wide range of logarithmically spaced scales, repeating RANSAC at least 30 times at each level, in order to identify the scale at which the Variance of the estimated model Parameters (VoP) is minimized.

A notable disadvantage of this approach is high computational cost: even with a modest granularity of 100 scales, this would require running RANSAC about 3000 times. One must also consider that running RANSAC with too small a scale imposes a near-zero inlier ratio, which requires an exponentially larger number of samples for the adaptive convergence criterion. This problem can be partially avoided by using an artificial limit on the number of iterations, although such a limit might prevent the true structure from being found if the outlier ratio is high.

Another more subtle problem is a dependence on model parameterization, because the algorithm assumes that the VoP is indicative of 'structural variation' of the model. Variance is completely meaningless for over-parameterized models (such as homogeneous entities), and even after projecting into a minimal parameterization (e.g., by performing a homogeneous division), variance is still not an accurate reflection of structural variation. For example, if one projects the homogeneous equation of a line into the familiar form of $y = mx + b$, one is likely to observe extremely high variance in the $b$ parameter for near-vertical lines, which is much greater than the variance would be for a set of nearly horizontal lines of equal angular variation. Thus, in order to obtain good results for any particular problem, one may need to spend a great deal of effort into finding a parameterization in which the VoP corresponds well to structural model variation. This alone makes it unsuitable as a generic estimation routine.

A related problem is that the algorithm requires comparing models to find the largest-scale model that is 'consistent' with the model at the scale that minimizes the VoP. In their implementation, model consistency is assessed by using the Frobenius norm of model parameters with some unspecified threshold, but the Frobenius norm of model parameters is generally not an accurate measure of 'structural difference' between models. Furthermore, choosing this threshold automatically is implicitly related to the noise level, and is arguably no simpler than choosing the original RANSAC threshold.

Lastly, the scale cannot be estimated more finely than the search discretization, and while it is generally true that the low-

est variance occurs around the true scale, this is not always the case. For example, if one is fitting lines to point data in $\mathbb{R}^2$, and there are two large outliers outside of a data set, then any threshold large enough to connect these two outliers will consistently result in the bad line connecting those outliers. Also, whenever there is a low outlier ratio, the scale encompassing all data points will be preferred over the true scale.

A more recent approach in the same spirit as StaRSaC is RECON (Raguram and Frahm, 2011), which also attempts to determine scale based on the recognition that model variance is low around around the true scale, but with one major difference: rather than explicitly looking for low variance in the model parameters, RECON looks for models with low variance in the sort-order of residuals (or fitting errors).

RECON forms model hypothesis from randomly selected minimal subsets until $K \geq 3$ models with mutually $\alpha$-consistent residual sets have been found. The $\alpha$-consistency test searches for the smallest $n$-value such that the data points associated with the $n$ smallest residuals have more than $\alpha^2$ percent overlap, with $\alpha = 0.95$. It is assumed that this $n$-value represents the separation between inliers and outliers. Although this test is statistically inspired, there are a number of practically occurring situations in which it fails.

First, it rests heavily on the implicit assumption that the fitting errors for inliers between any two good models will occur in random order. However, if one compares two *identical* models, then the fitting errors must have the exact same sort-order, and thus $\alpha$-consistency would pass at any value, such as the minimal value of $n = 1$, meaning that none of the inliers are detected.

In the presence of noise, it is unlikely to find two identical models, but the sort-order can still be expected to be similar for similar models. This is mostly a problem in the fourth step of RECON, which calls for making $M = 30$ over-determined model estimates from outlier free data, and than taking the minimal $n$-value that passes the $\alpha$-consistency test between all pairs. Because these estimates are over-determined and outlier free, it should be expected that *some* of these models are very accurate and hence very similar, and hence it would not be surprising if there were some pair of models with a very similar sort order, leading to a greatly under-estimated $n$-value. Because RECON then re-estimates the final model from this minimal number of points, it would destroy the model estimate.

Another problem is that, for very small values of $n$, the $\alpha$-consistency test can easily pass for inconsistent models by pure chance. For example, consider a line fitting problem with two perpendicular intersecting line models that both have the smallest fitting error to the same point nearby their intersection. In this case, the normalized overlap $\theta_1^{i,j} = 1$, and thus the two models will be deemed as $\alpha$-consistent, despite that these models are not at all consistent in their support regions. Although the individual probability of this for a single trial is low, given a sufficiently large number of $\alpha$-consistency trials, the probability of this problem occurring at least once becomes very large.

This problem can occur for arbitrarily large $n$ values, and may be exacerbated by the distribution of data points, because the only condition is that two inconsistent models happen to share some of their lowest residuals in common order (e.g., the models 'pivot' around a similar point in the data). Thus, if the data distribution inherently supports a region of common points that are likely to be shared by many different models (such as a bow-tie distribution for line fitting), then the same problem may be common for larger values of $n$. In the case of F-matrix estimation, a large number of points on a common plane could create this problem, even if the data also contains a significant number of off-plane points as well.

The runtime performance of RECON can be prohibitive, because for each RECON hypothesis, one must test for $\alpha$-consistency with all prior RECON hypothesis – and each test for consistency requires a brute force search through the residuals at all possible scales. Thus, despite that the overall number of samples is low, this high time complexity coupled with the large number of consistency checks required to find a set of mutually consistent models can quickly result in excessive runtime for low outlier ratios.

RECON also has difficulty with data sets that may contain multiple structures, because it returns the first significant structure that is found, which is not necessarily the most dominant structure in the data.

In the context of multiple model fitting, a number of methods have been proposed that also incorporate automatic scale selection (Toldo and Fusiello, 2009; Chin et al., 2009; Wang et al., 2012). However, Wang et al. (2012) have already developed ASSC and ASKC which are optimized for single-model estimation, Chin et al. (2009) effectively generalizes the principle used by RECON to the multiple model fitting problem, and the method of Toldo and Fusiello (2009) does not work for single-model estimation problems. Thus, we will not consider these more complex and computationally intensive multiple-model fitting methods further.

To summarize, it is hard to justify the greatly increased computational cost or reduction in reliability that is associated with using any of the aforementioned RANSAC variations that incorporate automatic scale selection, especially for computer vision problems, where the errors are generally measured in image space, and one can often assume a sub-optimal threshold that works 'acceptably well' based on the assumption that image correspondence error is on the order of a pixel or two, as is done with current state of the art Structure from Motion (SfM) systems like Bundler (Snavely et al., 2006, 2008).

This is not always the case, as one might be using a sub-pixel matching algorithm for low-baseline pairs leading to sub-pixel errors, or one might be performing wide-baseline matching with multi-scale features that have significantly larger errors, or tracking points across multiple frames resulting in the accumulation of errors, or dealing with images of varying sizes and qualities leading to unpredictable error levels. Thus, automatic scale estimation is still preferable, if it could be done reliably and efficiently.

Our recognition is that it is usually not difficult to specify a conservative maximum scale, and doing so permits the development of a new approach to the scale estimation problem without the large sacrifices in efficiency or reliability that are associated with completely unbiased searching through scale space. This is the motivation behind the proposed Simultaneous Fitting and Scale Estimation (SIMFIT) algorithm.

Like the original RANSAC algorithm, SIMFIT is simple to implement, is applicable to arbitrarily-high dimensional data, has no specific breakdown point, is independent of model parameterization, and uses the same statistical convergence criterion to adapt the number of iterations. It does not require any additional parameters, and does not significantly increase computational cost or reduce reliability. Furthermore, SIMFIT is designed to be fully general, and not just limited to computer vision problems.

We begin by introducing the theoretical background for classifying inliers (Section 2) and estimating scale from the residuals or fitting errors (Section 3). We then introduce the SIMFIT algorithm (Section 4), clarify the parameters of algorithms compared (Section 5) and present our experimental results (Section 6), starting with a validation of the assumed noise distribution (Section 6.1), followed by an empirical comparison of accuracy and performance on line fitting and homography estimation (Section 6.2) and finally a real-data experiment with fundamental matrix estimation (Section 6.3). Our results show that SIMFIT produces a model estimate with greater likelihood, more accurately estimated scale, and lower computational cost.

## 2 Classifying Inliers

For some implicit $p$-dimensional model defined by parameters $\theta \in \mathbb{R}^p$, let the function $f(\mathbf{x}|\theta) : \mathbb{R}^n \to \mathbb{R}^r$ be a mapping from data measurements to residual errors, where $n$ is the dimension of the measurement space and $r$ is the number of residual errors per datum. Thus, for a set of $N$ data measurements $\mathbf{x}_i \in \mathbb{R}^n, i = 1 \dots N$, the function $f(\mathbf{x}_i|\theta) = \mathbf{d}_i$ maps the $i$th datum to $\mathbf{d}_i$, a vector of *residual errors* associated with the datum. We call $||\mathbf{d}_i||^2$ the squared *fitting error* of $\mathbf{x}_i$ with respect to the model $\theta$, which is equal to zero only when $\mathbf{x}_i$ is perfectly consistent with $\theta$.

When fitting an $m$-dimensional surface in an $n$-dimensional space there are $k = n - m$ degrees of freedom in defining a surface normal, called the *codimension* (Kanatani, 1996). If measurement noise is independent and normally distributed with standard deviation $\sigma$, and given a reasonable model estimate $\hat{\theta} \approx \theta$, then residual errors in the codimension will be distributed approximately the same as measurement errors (normally). Thus, squared fitting errors will be distributed according to a scaled $\chi^2$-distribution with $k$ degrees of freedom (Dyer, 1973).

Once the scale $\sigma$ is known, a threshold may be calculated as $\tau^2 = \sigma^2 F_k^{-1}(\alpha)$ (Hartley and Zisserman, 2004, p.119), where $\alpha$ is the desired percentile (e.g., $\alpha = 0.95$) and $F_k^{-1}$ is the standard inverse cumulative $\chi_k^2$-distribution function. Given an estimated model $\hat{\theta}$ and threshold $\tau$, a datum may then be classified as an *inlier* when the squared fitting error is below the threshold; that is, $||\mathbf{d}_i||^2 < \tau^2$.

## 3 Robust Scale Estimators

Given an existing model estimate $\hat{\theta}$, a robust scale estimator attempts to estimate the true scale of measurement noise from the distribution of residuals or fitting errors relative to the model estimate. The maximum likelihood (ML) estimate of $\sigma$ is simply given by the sample standard deviation from the combined set of residuals. In the special case where the number of $\chi^2$ degrees of freedom is equal to the number of residuals per datum ($r = k$), this can be written in terms of the fitting errors as

$$\hat{\sigma}_{ML} = \sqrt{\frac{1}{Nk}\sum_{j=1}^{Nk} r_j^2} = \sqrt{\frac{1}{Nk}\sum_{i=1}^{N} ||\mathbf{d}_i||^2}, \tag{1}$$

where $r_j$ is the $j$th residual out of the combined set. However, it is well known that the ML estimate is not robust to outliers, and we expect that the data *does* contain outliers. A robust alternative comes from the median squared residual (Rousseeuw, 1987; Torr and Murray, 1997). Assuming that residual errors are distributed as $R \sim \mathcal{N}(0, \sigma^2)$, we have

$$0.5 = P(R^2 < \operatorname{med} R^2) = P(|R| < \operatorname{med}|R|) \tag{2}$$

$$= P(|Z| < (\operatorname{med}|R|)/\sigma), \tag{3}$$

where $Z = R/\sigma$ is a standard normal random variable (RV). Because the distribution of $Z$ is symmetric, the above is equivalent to

$$0.75 = P(Z < (\operatorname{med}|R|)/\sigma), \tag{4}$$

which implies

$$\Phi^{-1}(0.75) = (\operatorname{med}|R|)/\sigma \tag{5}$$

$$\sigma = \sqrt{\operatorname{med}|R|}/\Phi^{-1}(0.75), \tag{6}$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution. Therefore, a robust and asymptotically consistent estimator for $\sigma$ from the residuals is given by

$$\hat{\sigma}_{MED} = \frac{\operatorname{med}_j |r_j|}{\Phi^{-1}(0.75)} = \frac{\sqrt{\operatorname{med}_j r_j^2}}{\Phi^{-1}(0.75)}. \tag{7}$$

In practice, when $k > 1$, one does not always have a residual vector, and it is tempting to use (7) to estimate $\sigma$ from the fitting errors instead; however, this would be incorrect because the squared fitting errors have a non-central $\chi^2$ distribution. The median absolute deviation (MAD) is often used to compensate for this non-centrality, but this is not an asymptotically consistent estimator.

The correct estimator can be derived in the same fashion as (7). Specifically, if $D \sim \mathcal{X}^2(\sigma, k)$, then from the definition of the median we have

$$0.5 = P(D < \operatorname{med} D), \tag{8}$$

which implies

$$0.5 = F_k(\operatorname{med} D|\sigma) \tag{9}$$

$$\operatorname{med} D = \sigma^2 F_k^{-1}(0.5) \tag{10}$$

$$\sigma = \sqrt{(\operatorname{med} D)/F_k^{-1}(0.5)}. \tag{11}$$

Thus, a robust and asymptotically consistent estimator for $\sigma$, analogous to (7) but computed from the fitting errors and hence valid for all cases, is given by

$$\hat{\sigma}'_{MED} = \sqrt{\frac{\text{med}_i \, ||\mathbf{d}_i||^2}{F_k^{-1}(0.5)}}. \tag{12}$$

It should be noted that these median based estimators have 50% breakdown points, which is ineffective for most previous scale estimation algorithms where the breakdown point of the scale estimator induces an equivalent breakdown point in the overall estimation routine (Wang and Suter, 2004a; Fan and Pylvänäinen, 2008; Wang et al., 2010; Raguram and Frahm, 2011).

This has led to the development of scale estimators with increased tolerance to outliers, such as the Compressed Histogram (CH) (Yu et al., 1994), the $k$-th order statistic (Lee et al., 1998; Bab-Hadiashar and Suter, 1999), the weighted median absolute deviation (WMAD) (Fan and Pylvänäinen, 2008), Two-Step Scale Estimator (TSSE) (Wang and Suter, 2004a), IKOSE (Wang et al., 2012) and others.

However, the breakdown point of the scale estimator is not a significant concern for SIMFIT, becuase the scale estimate is only used to obtain an over-estimate anyway. Thus, we will prefer $\hat{\sigma}'_{MED}$ for its simplicity and reliability.

## 4 Simultaneous Fitting and Scale Estimation (SIMFIT)

The sensitivity to threshold choice $\tau$ in RANSAC is revealed by the fact that, as $\tau \to \infty$, all constraints on the estimated model vanish, giving a purely random result. In contrast, we notice that MSAC (Torr and Zisserman, 2000), a modification of RANSAC that minimizes a robust M-estimator (Huber, 1981), becomes equivalent to the method of least absolute deviations (LAD) (Branham, 1982) as $\tau \to \infty$.

LAD is already a fairly robust method, and by using any $\tau < \infty$, one may obtain far more robust results without a specific breakdown point. Thus, MSAC is quite robust to overestimated scales, and this is the core concept we exploit in the algorithm outline below:

1. Starting from any initial overestimate of $\sigma$, the corresponding optimal threshold $\tau$ may be derived, and used to estimate a model with associated inliers using MSAC.
2. From the residuals of the inlier set, a robust estimate of $\sigma$ may be computed using $\hat{\sigma}'_{MED}$. Because it was estimated from a more restricted set of inliers, the newly estimated $\sigma$ will usually be less than the previous.
3. If there is no significant change in the estimate of $\sigma$, then all the outliers must have been removed, and hence the model, inliers, and scale should all be accurate. Otherwise, one may repeat MSAC from step 1 using the newly reduced estimate of $\sigma$.

To clarify the algorithm details, we give pseudo-code in Algorithm 1, and proceed here with some analysis. First, the initial estimate of $\sigma$ is used to calculate a corresponding overestimate of $\tau$ (line 2), and all the data points are added to the potential inlier set (line 5).

**Algorithm 1** SIMFIT

**Require:** *data* is a set of measurements, $\sigma_{MAX} > \sigma$ is an overestimate of the true inlier noise, $\epsilon \geq 0$ is the minimum error tolerance desired in the scale estimate, $\alpha \in (0, 1)$ is the desired percentage of inliers to capture with a threshold on fitting error (e.g., 0.99).

**Ensure:** $(\hat{\theta}, \hat{\sigma}, inliers)$ is the final estimated model, with corresponding estimate of the noise scale and inlier set.

1: $\hat{\sigma} \leftarrow \sigma_{MAX}$
2: $\tau^2 \leftarrow \hat{\sigma}^2 F_k^{-1}(\alpha)$
3: $hashStates \leftarrow \{\}$
4: $inliers \leftarrow \{\}$
5: **for** $i = 1$ **to** $\#data$ **do**
6: Add $i$ to $inliers$
7: **end for**
8: **repeat**
9: $\hat{\theta} \leftarrow$MSAC($data$, $inliers$, $\tau$)
10: **for all** $i \in inliers$ **do**
11: **if** $||\mathbf{d}_i||^2 > \tau^2$ **then**
12: Remove $i$ from $inliers$
13: **end if**
14: **end for**
15: $\hat{\sigma}_{prev} \leftarrow \hat{\sigma}$
16: $\hat{\sigma} \leftarrow \hat{\sigma}'_{MED}(inliers)$
17: $\tau^2 \leftarrow \hat{\sigma}^2 F_k^{-1}(\alpha)$
18: $h \leftarrow$hash($inliers$, $\hat{\sigma}$)
19: **if** $hashStates$ contains $h$ **then**
20: $cycleDetected \leftarrow$ **true**
21: **else**
22: Add $h$ to $hashStates$
23: **end if**
24: **until** $|\hat{\sigma} - \hat{\sigma}_{prev}| \leq \epsilon$ **or** $cycleDetected$
25: **repeat**
26: Improve estimate of $\hat{\theta}$ from all $inliers$
27: $modified \leftarrow$ **false**
28: **for all** $\mathbf{x}_i \in data, i \notin inliers$ **do**
29: **if** $||\mathbf{d}_i||^2 < \tau^2$ **then**
30: Add $i$ to $inliers$
31: $modified \leftarrow$ **true**
32: **end if**
33: **end for**
34: **until** $modified =$ **false**
35: $\hat{\sigma} \leftarrow \hat{\sigma}'_{MED}$(newInliers)

On each iteration, MSAC is used to compute a robust estimate of the model $\theta$ (line 9) from within the potential inlier set (our modified version of MSAC that works with a shrinking inlier set is given in Algorithm 2). The set of inliers is reduced (line 10) and used to compute a new robust estimate of scale (line 16) and associated threshold (line 17).

In general, each new estimate of scale will be lower than the previous until convergence. However, this is not guaranteed, and in some very rare cases a cycle might be entered. Therefore, we perform explicit cycle prevention by computing the MurmurHash of the inlier indices and current scale estimate (rounded to nearest integer), and break out of the loop if a repeated state would be entered (line 24).

Normally, convergence is detected when the reduction in $\sigma$ becomes insignificant, as detected by a difference less than some threshold $\epsilon$ (line 24). However, we note that one may ignore this parameter by setting $\epsilon = 0$ here, which merely delays convergence until no further improvement is possible.

Additionally, one may add some optional convergence criteria to improve best and worst case performance: (a) If the

found solution uses nearly all of the potential inliers (i.e., if the number of inliers reduced from the current iteration is an insignificant fraction); (b) If the found solution uses such a small number of inliers that further reduction of the inlier set would be pointless (i.e., the size of the current inlier set is less than 2 times the minimal number of points necessary to define a model).

**Algorithm 2** MSAC, modified for shrinking inlier set

**Require:** *data* is a set of measurements, *inliers* is a set of data indices that are not known to be outliers, $\tau > 0$ is the inlier threshold, $pFail \in (0, 1)$ is the accepted probability of failure (e.g., $1 \times 10^{-4}$), and *pickSize* is the number of measurements to use in each random selection (determined by model estimator).

**Ensure:** The model $\hat{\theta}$ has been estimated from only inliers with probability $1 - pFail$.

1: $minCost \leftarrow \infty$
2: $iters \leftarrow 0$
3: **repeat**
4: $pickSet \leftarrow$ Select *pickSize* unique indices from *inliers*
5: Estimate $\hat{\theta}_{test}$ from *pickSet*
6: $cost \leftarrow 0$
7: $count \leftarrow 0$
8: **for all** $i \in inliers$ **do**
9: Compute $||\mathbf{d}_i||$ using $\hat{\theta}$
10: **if** $||\mathbf{d}_i|| < \tau$ **then**
11: $cost \leftarrow cost + ||\mathbf{d}_i||$
12: $count \leftarrow count + 1$
13: **else**
14: $cost \leftarrow cost + \tau$
15: **end if**
16: **end for**
17: **if** $cost < minCost$ **then**
18: $minCost \leftarrow cost$
19: $\hat{\theta} \leftarrow \hat{\theta}_{test}$
20: $needIters \leftarrow \log(pFail)/\log\left(1 - (count/\#inliers)^{pickSize}\right)$
21: **end if**
22: $iters \leftarrow iters + 1$
23: **until** $iters \geq needIters$

In most cases, we do not expect to need more than 1-3 iterations of MSAC to converge to the correct scale. Moreover, because MSAC is run within a reduced inlier set (similar to LO-RANSAC (Chum et al., 2003)), subsequent runs of MSAC become computationally trivial, as the inlier ratio will be near to 1, requiring only a few random samples to meet the statistical convergence criterion of Fischler and Bolles (1981).

After convergence to the proper scale, we transition into a final (optional) model refinement stage (line 25), which we refer to as the *model-shift* procedure, because it is actually a generalization of the well-known mean-shift procedure (Comaniciu and Meer, 2002), where the threshold is effectively the mean-shift bandwidth with a uniform kernel, and we generalize the sample mean from mean-shift with the over-determined estimate of the model. The only actual difference from mean-shift is that we *only* allow inliers to be added (and not removed) from the potential inlier set, which guarantees convergence by preventing cycles.

SIMFIT is usually quite robust to the choice of $\sigma_{MAX}$. For example, if one chooses $\sigma_{MAX} = \infty$, then the first iteration would reduce $\sigma_{MAX}$ down to $\hat{\sigma}'_{MED}$ from an all-data fit. This is often sufficient to converge to the proper scale, although when the outlier points come from some distribution that also has finite variance, then the all-data fit may yield a stable model that is a false attractor. Thus, one should choose $\sigma_{MAX} < \hat{\sigma}'_{MED}$ if possible.

## 5 Algorithms Compared

In this section we identify previous algorithms that we compare to SIMFIT in our experimental results for their ability to do robust estimation with simultaneous scale detection. We also clarify the choice of free parameters and algorithm details when necessary. In general, we set $\tau$ so as to capture $\alpha = 0.99$ percent of inliers, and we let the number of samples for RANSAC/MSAC be determined adaptively with $pFail = 1 \times 10^{-3}$, and a maximum of 10000.

### 5.1 RANSAC

Although RANSAC does not include scale estimation, we use RANSAC with an optimally derived threshold based on the true $\sigma$ for the purposes of performance comparison.

### 5.2 SIMFIT

Because we expect image noise on the order of a pixel or so, we use a conservative over-estimate of $\sigma_{MAX} = 15$. We use the same value in synthetic tests. We use three conditions for early-termination of the main loop: (a) change in $\sigma$ less than $\epsilon = 0.5$, (b) reduction in the inlier set is less than 1% from the previous iteration, (c) size of the inlier set is reduced to less than 2 times the minimal number of points to define a model.

### 5.3 LMS+MSAC

The first comparative algorithm is the procedure of Torr and Murray (1997), where LMS is used to obtain an initial estimate of the model, followed by robust scale estimation using $\hat{\sigma}_{MED}$, and finally RANSAC with a threshold based on $\hat{\sigma}_{MED}$. We have upgraded this method by replacing RANSAC with MSAC from Torr and Zisserman (2000) because it is strictly superior to RANSAC in this context. LMS is implemented (as per usual) by random sampling, and the number of iterations is determined based upon the assumption that the data may contain up to 50% outliers (because this is the breakdown point of LMS).

### 5.4 ASSC and ASKC

The second algorithm that merits comparison is ASSC (Wang and Suter, 2004a), which modifies the RANSAC criterion to maximize the number of inliers divided by a robust scale estimate, as well as the newly developed ASKC (Wang et al., 2010). ASSC uses adaptive sampling, but ASKC uses a fixed number of samples, and we chose $M = 1000$ samples.

Both ASSC and ASKC estimate scale using TSSE, which uses a kernel density estimate (KDE) of the residual error,

where the bandwidth for the KDE is chosen automatically using a rule of thumb multiplied based on an initial scale estimate, and then multipled by an unspecified tuning parameter $c_h \in (0, 1)$ to compensate for oversmoothing.

For the initial scale estimate, we use the $k$-th order static with $k = 0.2$, as described in Wang et al. (2010). We use the default value of $c_h = 1$ because we are testing generic performance and do not permit tuning the algorithms for particular noise distributions.

### 5.5 RANSAC EIS M

Several comparable algorithms were proposed in Fan and Pylvänäinen (2008): RANSAC-MAD was essentially the same as ASSC but used the median absolute deviation (MAD) to estimate scale instead of TSSC, RANSAC-EIS modified the objective by calculating a weight vector using Ensemble Inlier Sets (EIS) and then using WMAD instead of MAD, and finally RANSAC-EIS-Metropolis incorporated weighted sampling. The latter was found to be the superior version, so we only compare against this one. We note that their algorithm calls for some number of unspecified fixed iterations, so we use the same probabilistic argument from RANSAC and LMS to calculate the number of required iterations assuming there are 50% outliers.

### 5.6 STARSAC

STARSAC requires a minimum and maximum scale, we used $\sigma_{MIN} = 1 \times 10^{-5}$ and $\sigma_{MAX} = 1000$, and tested at 20 scales logarithmically spaced within this range, performing the recommended 30 runs of RANSAC at each scale. Due to the large number of RANSAC iterations, and because the required number of RANSAC samples grows exponentially for underestimated scales, we found it computationally necessary to impose a maximum of 100 samples for each run of RANSAC.

When dealing with homogeneous models, the set of model parameters was taken as the set of real parameters after performing homogeneous division. Two models were deemed 'consistent' if the maximum relative difference between model parameters was less than 0.2.

### 5.7 RECON

The version of RECON that we use has a few improvements over the algorithm described in Raguram and Frahm (2011). First, we clarify that the termination condition of finding 'K mutually $\alpha$-consistent models' requires clustering models into mutually $\alpha$-consistent groups. These clusters are efficiently maintained using the union-find structure.

Because the initial $\alpha$-consistency test will always pass for inconsistent models at some large scale encompassing all (or most of) data, RECON explicitly tests for $\hat{\sigma}_{MED} < \sigma_{MAX}$, or alternatively tests the KS-test for distribution equality. We opted to use the scale-based test in our experiments because it is faster, more comparable to SIMFIT, and because the KS-test will often accept the null-hypothesis of distribution equality for an all-data fit when the outlier distribution has finite variance, inducing a breakdown point as a set of $\alpha$-consistent models are more likely to be found at maximum scale for low outlier ratios.

However, when the outlier ratio is less than 50%, the robustness of the $\hat{\sigma}_{MED}$ estimator may cause a low value of $\sigma$ to be estimated even when $n$ is large enough to be an all-data fit. This issue is corrected by instead testing against the sigma implied by the residual error of the $n$th residual. That is, $\sqrt{r_n^2 / F_k^{-1}(0.5)} < \sigma_{MAX}$.

Another issue is the assumption that residuals will always occur in random sort order. As discussed in the introduction, this assumption is invalid for over-determined models, and can lead to $\alpha$-consistency check passing at an incorrect small $n$-value. This can be problematic in the fourth step of the original algorithm, which takes the minimum $n$ value across all pairs between the $M$ over-determined models. This issue is corrected by recomputing the scale from the final model using $\hat{\sigma}_{MED}$, and then reclassifying inliers according to a threshold derived from this final scale estimate.

We use $\sigma_{MAX} = 15$ in all experiments. Additionally, we use $K = 3$ and $\alpha = 0.95$ as recommended by the authors. The authors recommend $M = 30$, but we found this to be computationally limiting, and reduced it to $M = 5$. To generate the final non-minimal models, we always use a sample size equal to twice the number of minimal points. Due to problems with the $\alpha$-consistency test for small $n$-values, we only consider $n > 10$.

## 6 Experimental Results

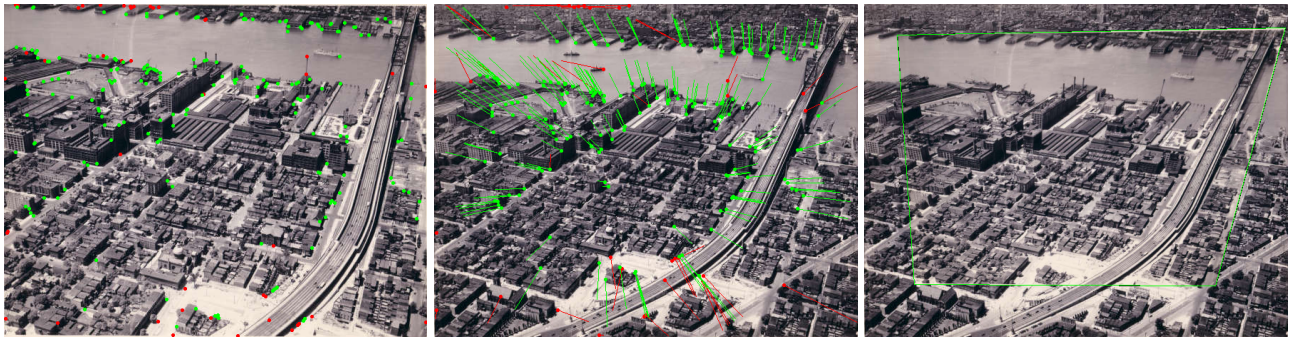
We begin with some experiments to validate that it is reasonable to assume squared fitting errors will be $\chi_k^2$-distributed (Section 6.1). Then we compare the accuracy, reliability and performance of SIMFIT against previous methods using a synthetic line fitting problem as well as homography estimation (Section 6.2), and finally on real data for fundamental matrix estimation (Section 6.3).

### 6.1 Analysis of Correspondence Noise Distribution

A first assumption used by all algorithms is that the inlier noise distribution is normal, which implies squared fitting errors of an ideal model will be $\chi_k^2$-distributed. However, to our knowledge, this assumption has never been validated in practice. Moreover, it is unclear how sensitive an algorithm would be to departures in normality of the measurement noise. We explore both of these issues here.

First, we consider robust estimation of a homography from a set of sparse correspondences for image registration. Because a homography perfectly describes the mapping of a planar surface between any two arbitrary viewpoints, it is a good model for registering aerial images where the ground surface is well-approximated by a plane. Because each corresponding point provides 2 constraints, we expect that squared fitting errors will be $\chi^2$-distributed with $k = 2$ degrees of freedom.

We obtained two aerial photographs (Fig. 1, left and middle) taken from slightly different positions at different times



(a) (b) (c)

Fig. 1. Example of a robust homography fit used to register two aerial photographs, courtesy of the Hagley Museum and Library (Hag, 1935a,b). Correspondences were found automatically by local patch similarity and then inliers identified using SIMFIT. (a) first image with feature points, green points are inliers to the found homography; (b) second image with corresponding feature points, green points are inliers to the found homography; (c) first image warped into the reference frame of the second image using the estimated homography.

of the day (as evidenced by the boats that have moved) and then automatically computed a set of typical correspondences by matching Harris feature points. These correspondences were filtered by SIMFIT to determine a set of inliers (indicated by the green points and lines in Fig. 1), along with the scale of inlier noise and an estimate of the homography, which was then used to register the first image into the frame of the second image for visual verification (Fig. 1, right).

There were a total of 244 potential correspondences found by the feature matcher. We used an initial over-estimate of $\sigma = 15$ pixels for SIMFIT. The first run of MSAC found 231 inliers and reduced the scale estimate to $\sigma = 1.21$ pixels. The second run of MSAC found 189 inliers and further reduced the scale estimate to 1.09 pixels. This resulted in convergence because our threshold is set at $\epsilon = 0.5$, causing SIMFIT to transition into the nonlinear model-shifting mode where it increased the potential inlier set to 199 correspondences and the final threshold was $\tau^2 = 3.31$.

We computed the KDE of the distribution of residual errors using the rule-of-thumb bandwidth (Silverman, 1986, p.48) and compared it to the PDF of the normal distribution using the found scale (Fig. 2, left), observing good visual agreement (note that the KDE is a bit rough given that there are only 199 samples). We then computed the ECDF of the fitting errors and compared them to a $\chi^2_2$ distribution at the appropriate scale factor, and again observed good visual agreement (Fig. 2, right). Thus, we conclude that the assumption of normality is reasonable.

Of course, we do not expect that measurement errors will always be normally distributed for all types of problems. Therefore, we are curious to investigate how different noise distributions will effect the distribution of fitting errors. To this end, we generated some synthetic data sets where the measurement noise was uniformly distributed as well as distributed according to an infinite variance stable distribution (Nolan, 2011).

The stable distribution arises as a generalization of the central limit theorem: whereas the central limit theorem states that the sum of any independent and identically distributed (IID) random variable (RV) with *finite* variance converges to a normal distribution, the generalized central limit theorem (Nolan, 2011) states that the sum of *any* IID RV converges to a stable distribution. Thus, the normal distribution, Cauchy distribution, Levy (or Pearson) distribution, Landau distribution and Delta distribution can all be found as special cases of the stable distribution. Because of its theoretical generality, the stable distribution is often suggested as a model for non-normal data.

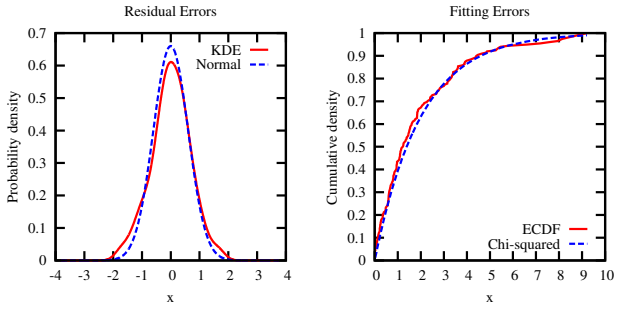



Fig. 2. Comparison between the empirical distribution and theoretical expectation after fitting a homography to a set of real image correspondences. Left: the kernel density estimate (KDE) of residual errors using the rule-of-thumb bandwidth compared to the PDF of a normal distribution. Right: the empirical cumulative distribution function (ECDF) of squared fitting errors compared to the CDF of a $\chi^2$-distribution.

When measurement noise is uniformly distributed, we see that the distribution of squared fitting errors is still quite well-approximated by the $\chi^2$ distribution (Fig. 3, left). In fact, even when measurement errors are distributed according to a stable distribution with $\mathcal{S}(\alpha = 1.5,\ \beta = 0,\ \gamma = 5,\ \delta = 0)$, which has infinite variance, we still observed fairly good visual agreement with the $\chi^2$ distribution (Fig. 3, right). This can be explained by the fact that any tail effects of the stable distribution are automatically chopped off and considered as outliers by the algorithm.

The assumption that fitting errors will be $\chi^2$-distributed is only relevant for picking an inlier threshold to capture the desired percentage of inliers, but this is an insensitive parameter that is just set with an arbitrary ballpark estimate anyway.

Thus, although the estimated scale may be off, minor divergences from the $\chi^2$-distribution are not otherwise relevant, and hence one does not generally need to worry about departures from normality in the residuals due to the correspondence detector or approximated error metrics.

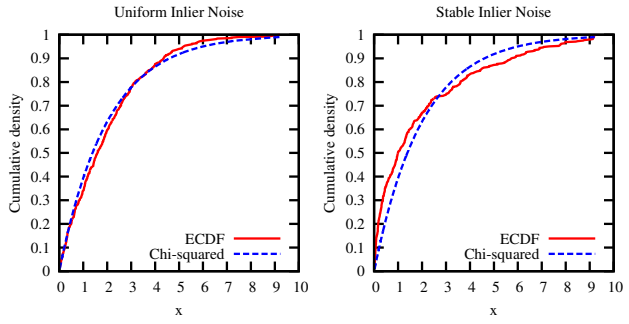



Fig. 3. Comparison between the ECDF of squared fitting errors and the predicted $\chi^2$-distribution in the robust estimation of a homography from synthetic data that intentionally violates the assumptions of normality. Left: using uniformly distributed inlier noise of $(-5, 5)$ pixels, the $\chi^2_2$ approximation still works well; Right: even when the inlier noise comes from a stable distribution ($\alpha = 1.5$, $\beta = 0$, $\gamma = 5$, $\delta = 0$), which has infinite variance, the $\chi^2$ approximation is still not bad.

## 6.2 Experiments on Synthetic Data

Our experiments are designed to test all algorithms starting from outlier-free data up to the point of failure under extreme contamination from outliers and noise. The first test is 2D line fitting. For each trial, outliers were uniformly distributed in a $500 \times 500$ region, and the 'true line' was chosen by joining two random points from the outlier distribution. A random noise scale was chosen $\sigma \in (1, 10)$, and inliers were generated by choosing a random $t \in (0, 1)$ to interpolate between the two endpoints of the line (where it intersects the bounding box of the $500 \times 500$ region), and then adding normally distributed noise with standard deviation $\sigma$ to each coordinate.

Because a line is one-dimensional, the codimension of fitting a line to $n$-dimensional points is $k = n - 1$. Thus, the squared fitting error is distributed according to a $\chi^2$-distribution with $k = 1$ degrees of freedom.

Our second test was estimation of a homography from 2D correspondences. The true homography was chosen as a random rotation matrix with $\theta \in (0, 2\pi)$, although we used a search space of all $3 \times 3$ homographies. Inlier correspondences were generated by choosing a random point in the $500 \times 500$ region for the first point, which was transformed by the true homography and then normally distributed noise was added to each coordinate of the second point.

For both tests, we vary the fraction of outliers from $R = 0 \ldots 0.9$; for each $R$-level, we generate 100 random data sets consisting of 1000 points each, and plot the median of several performance statistics for each algorithm:

**Found inliers** This is the number of inliers found by the algorithm; ideally, it should be roughly $1000(1 - R)$ because there are 1000 measurements.

**Found scale** This is the estimated scale divided by the true scale, so the ideal value is 1. Note that each data set has a random scale $\sigma \in (1, 10)$.

**Error** This is an objective measure of the model error, calculated as the sum of squared fitting errors from all of the *true* inliers divided by the sum of squared errors from the true model. Thus, the ideal value is 1.

**#Minimal fits** This records the total number of random models that were evaluated by the algorithm.

**Runtime(sec)** The total runtime of the algorithm in seconds (tested on a Lenovo X220 laptop with Core i7-2620M processor).

Some visual examples from the line fitting problem are shown in Fig. 4, where the capacity of SIMFIT to routinely extract an accurate line from heavily contaminated data without any prior knowledge of the scale of inlier noise can be observed.

We also show a proof of concept of how SIMFIT can be used to extract multiple models with different noise scales from the same data set (Fig. 4, c). This is done by repeatedly fitting a model with SIMFIT to extract the most dominant structure, removing all the found inliers, and then running SIMFIT again to extract the next most dominant structure in the remaining data, until the inlier count falls below some threshold. It remains to be seen how this fit-and-remove method compares with other dedicated multiple-model fitting algorithms that perform scale selection (Toldo and Fusiello, 2009; Chin et al., 2009; Wang et al., 2012).

### *6.2.1 Breakdown Point Analysis*

The breakdown point can be identified as the highest outlier ratio in which the algorithm succeeded in finding a reasonable fit, rather than breaking down to an all-data fit. For line estimation (Fig. 5, top left), LMS+MSAC has a breakdown point at 50% (as predicted), ASSC had a breakdown point at about 70%, RANSAC EIS M had a breakdown point at about 60%. The other algorithms did not break down, but at $R = 0.9$, RECON and ASKC began to drastically over-estimate scale (Fig. 5, top middle) and STARSAC had very high model error (Fig. 5, top right). Thus, we may conclude that these algorithms are on the verge of breaking down.

For homography estimation (Fig. 6, top left), ASCC broke down earlier at around 40%, and ASKC also brokedown at 70%. RECON also appears to breakdown at 70%, but this is an artificial breakdown point due to exceeding the 1 hour time limit that we set (after which we resort to an all-data fit). In theory, RECON should not have a breakdown point. At $R = 0.9$, SIMFIT was on the verge of breaking down, but this was usually corrected by the model shift procedure.

Overall, we conclude that SIMFIT is the most robust method to breakdown, being the only algorithm surveyed that did not break down for either problem (other than RANSAC with optimal threshold choice, which does not perform scale selection, and perhaps RECON, given infinite time).

### *6.2.2 Scale Estimation Analysis*

The estimated scale for line fitting is shown in (Fig. 5, top middle) and(Fig. 6, top middle) for homography estimation.

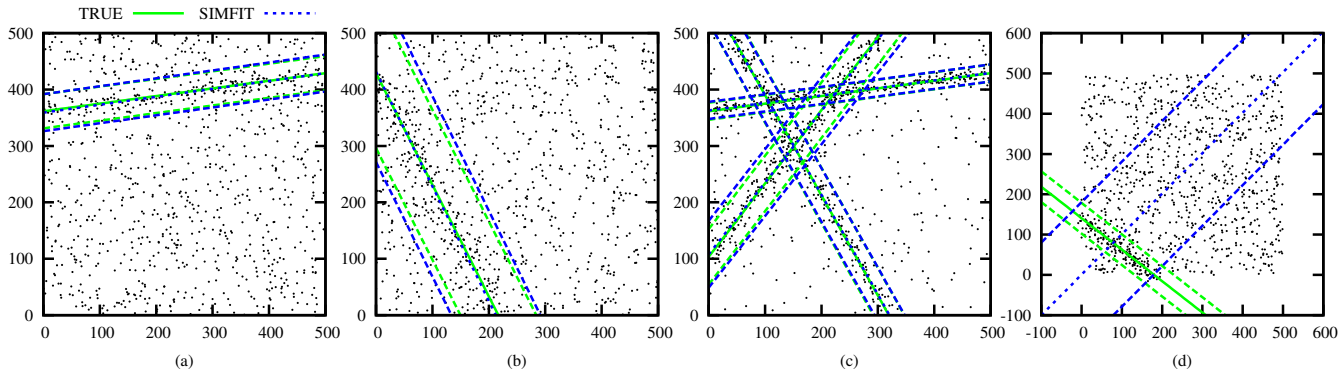



Fig. 4. Examples of the estimated SIMFIT line (blue) in comparison to the true line (green). The lines at $\pm 3\sigma$ are also shown as a visual representation of the estimated scale. (a) An accurate result at $R = 0.85$; (b) An accurate result at $R = 0.8$; (c) Example of multiple model extraction, with 25% inliers for each of the three models; (d) A failure case, where the outlier distribution creates a false attractor encompassing most of the data.

In the presence of outliers, we see that LMS+MSAC tends to over-estimate scale, with a gradually increasing over-estimate up until it reaches the breakdown point. This mirrors the performance of the underlying $\hat{\sigma}_{MED}$ estimator. RECON shows a similar trend because it uses the same estimator, but because the breakdown point is so much higher, this results in much more accurate estimates, generally between 1-2 times the correct scale.

RANSAC EIS M was specifically designed to compensate for the over-estimated scale of LMS+MSAC (Fan and Pylvänäinen, 2008), but we see that it instead under-estimates scale before reaching its breakdown point, after which it also over-estimates the scale.

On the other hand, STARSAC has a large degree of variance in the scale estimate. This is because the search granularity does not permit precise scale estimates, and because the check for model 'consistency' is based on a sub-optimal threshold choice. We observed greater variability in performance on the homography problem because the variance of model parameters becomes increasingly less representative of 'structural' model variation for higher dimensional models.

ASSC and ASKC tend to under-estimate scale (by about 50%) when they are operating well below the breakdown point for line estimation, and then transition into over-estimating scale as they near the breakdown point. ASKC is more sensitive to this under-estimated scale, detecting only about 50% of the inliers for low outlier ratios, despite finding a good model (Fig. 5, top right). This is because when the bandwidth is under-estimated, it changes the normalization factor of the KDE such that an under-estimated scale may have an equally high (or higher) kernel density at the origin of residual space. This problem did not occur for homography estimation (Fig. 6, top left), but only because the bandwidth was over-estimated, resulting in over-estimated scale estimates (Fig. 6, top middle).

It should be noted that these problems with ASSC/ASKC can be avoided by parameter tuning. We used the recommended values from Wang et al. (2010), but different results are achieved by changing the $k$-value in the $k$th order statistic, or the $c_h$ parameter, or using a different robust scale estimator. For example, using $\hat{\sigma}_{MED}$ as suggested in Wang and Suter (2004a), there was no difficulty in capturing all the inliers for low outlier ratios, but a breakdown point was introduced at around 30%. However, as evidenced by the tendency to over-estimate in one case (Fig. 6, top middle) and under-estimate in another (Fig. 5, top middle), there is no clear way to tune these parameters that works well for all cases.

Although SIMFIT began to slightly over-estimate scale for the line fitting problem at $R = 0.9$, we see that overall, it was the only algorithm that consistently and accurately found the true scale. The final model shift procedure performed negligible improvement for line fitting because the initial estimate was quite good; however, we see that for the higher dimensional problem of homography estimation where scale is slightly over-estimated initially, the model shift procedure consistently corrects this estimate to find the true value.

### *6.2.3 Model Error Analysis*

For the line estimation problem, the reconstructed model error (Fig. 5, top right) is fairly comparable for all methods up until they reach their breakdown points, being consistently in the range of 1-2 times the minimal error value. Nonetheless, SIMFIT performed the best, being the only algorithm that consistently found the minimum error at all outlier ratios (after model shifting).

For the higher dimensional homography estimation problem, the variation in model accuracy between algorithms is more pronounced (Fig. 6, top right) . First, we note that even RANSAC with optimal threshold tends to find a model with about twice the minimal error. ASSC often has more than 4-5 times the minimal error, and LMS+MSAC often had 3-4 times the minimal error. ASKC, STARSAC, RECON, and RANSAC EIS M were all capable of finding less than 2 times the minimal error, but this quickly increased as their breakdown points are neared.

Interestingly, the core SIMFIT estimation actually became more accurate as the outlier ratio increased, up until $R = 0.9$ where there was a significant increase in model error. Still, after model shifting, SIMFIT consistently found the minimum error at all outlier ratios.

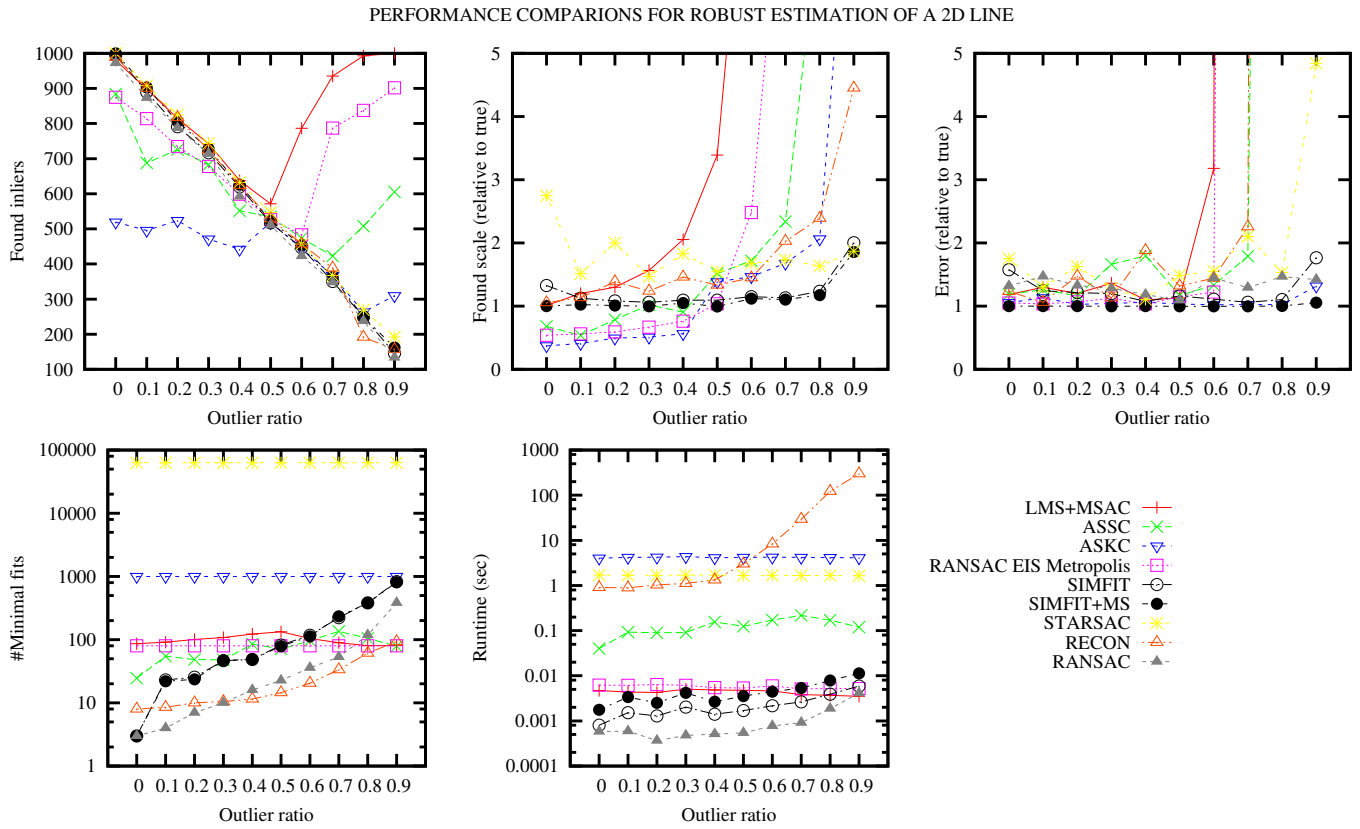



Fig. 5. Performance comparisons for the robust estimation of a 2D line with scale estimation. See text for a full description of the experiment.

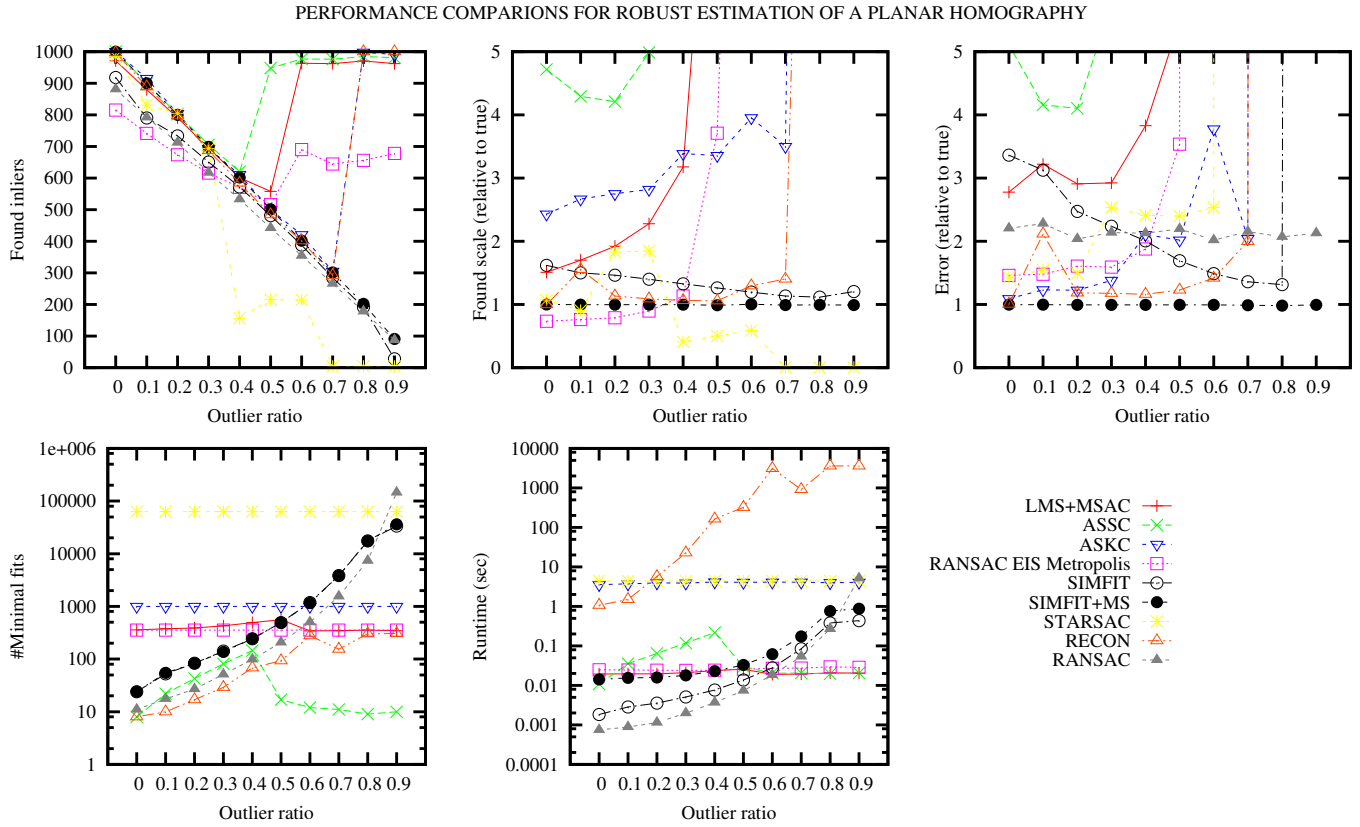



Fig. 6. Performance comparisons for the robust estimation of a 2D homography with scale estimation. See text for a full description of the experiment.

*6.2.4 Performance Analysis*

RANSAC EIS M, STARSAC, and ASKC do not attempt adaptive sampling, which means that runtime performance is independent of outlier ratio. Although MSAC does adaptive scaling, the overall performance of LMS+MSAC is dominated by the LMS phase which requires a fixed number of iterations, and because LMS breaks down for $R > 0.5$, MSAC is not able to scale beyond this point. As a result, the number of unique fits in LMS+MSAC is also effectively constant.

Lack of adaptive sampling is a notable disadvantage because choosing the optimal number of iterations in advance requires *a priori* knowledge the true outlier ratio. If too few iterations are chosen, a breakdown point may be induced. If too many iterations are chosen, performance for low outlier ratios will suffer

It is interesting to note that, despite requiring nearly 100 times as many samples, STARSAC had almost the same performance as ASKC (about 4 seconds) (Fig. 5, bottom right), due to the costly objective function of ASKC that requires using the KDE.

With ASSC, we see that adaptive sampling works up until $R = 0.4$ (the breakdown point), and then scales back down (Fig. 6, bottom middle), despite that the most samples are needed for high outlier ratios. This is because as soon as a model is found with an over-estimated scale that can over-classify an inlier set, it causes a reduction in the required number of iterations – despite that this over-estimated model may be incorrect. In other words, it introduces an artificial breakdown point, and this is likely why adaptive sampling was later abandoned in ASKC.

SIMFIT and RECON are the only methods that consistently succeed in adaptively scaling up the number of iterations. Using $K = 3$, RECON often requires fewer model hypothesis than SIMFIT or even RANSAC (at higher outlier ratios) – however, the large number of $\alpha$-consistency checks leads to high computational complexity, and runtime that is often several orders of magnitude larger. For example, on the homography estimation problem, the performance of SIMFIT ranges from 0.01 to 1 second, whereas RECON ranges from 1 second to more than 1 hour (and sometimes needed to be aborted due to time constraints) (Fig. 6, bottom right).

In contrast, we see that the performance of the core SIMFIT routine (without model shift) was very similar to RANSAC, being the only algorithm that remains in the same order of magnitude for all outlier ratios, with median performance about 2 times slower than RANSAC. For example, at $R = 0.5$, RANSAC took $7.42 \times 10^{-3}$ seconds, whereas SIMFIT took $1.37 \times 10^{-2}$ seconds. The final model shift procedure added roughly constant time, never more than 0.4 seconds.

## 6.3 Experiments on Real Data

Although we have so far demonstrated SIMFIT's capabilities on synthetic problem with higher outlier ratios and noise levels than typical vision problems, it should be stressed that SIMFIT is also competitive when applied to the lower outlier ratios and noise levels of typical correspondence data. We demonstrate this on another problem from computer vision: estimation of the fundamental matrix from image correspondences.

The fundamental matrix is a more general constraint on images than a homography because correspondences between *any* two images must obey the epipolar constraint, regardless of scene geometry. Given a set of corresponding image points $\tilde{\mathbf{x}}_i \leftrightarrow \tilde{\mathbf{x}}'_i$, both homogeneous points in $\mathbb{P}^2$, the epipolar constraint (Hartley and Zisserman, 2004) dictates that the fundamental matrix $\mathbf{F}$ should satisfy

$$\tilde{\mathbf{x}}_i'^{\mathsf{T}} \mathbf{F} \tilde{\mathbf{x}}_i = 0, \quad \forall i. \tag{13}$$

Geometrically, this constraint represents the fact that each point $\tilde{\mathbf{x}}_i$ in the first image defines an epipolar line $\mathbf{l}_i = \mathbf{F}\tilde{\mathbf{x}}_i$ in the second image, and the second point $\tilde{\mathbf{x}}'_i$ should lie on this epipolar line, so $\tilde{\mathbf{x}}'_i \cdot \mathbf{l}_i = 0$. We performed minimal estimation of the fundamental matrix from 6 points as described in Hartley and Zisserman (2004), and refined from over-determined point sets using bundle adjustment. Because the epipolar line constraint is a 1-dimensional constraint, we use $k = 1$ degrees of freedom for the $\chi^2$-distribution.

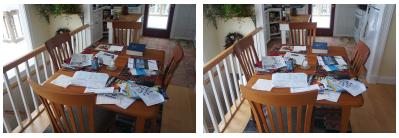
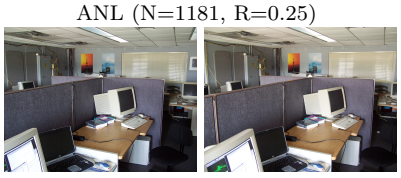
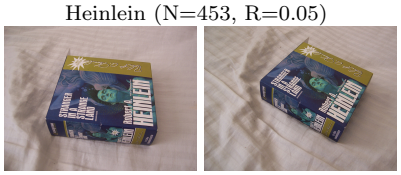
It is tempting to measure error as the squared distance from the right correspondence point $\tilde{\mathbf{x}}'_i$ to the epipolar line $\mathbf{l}_i$, as was done in Torr and Murray (1997). However, this introduces a bias by assuming that all of the noise is distributed on the measurements of $\tilde{\mathbf{x}}'_i$ rather than $\tilde{\mathbf{x}}_i$. Therefore, we prefer the maximum likelihood (ML) method, which is to simultaneously estimate the fundamental matrix and 3D structure points that minimize reprojection error.

We automatically generated correspondences by finding Harris corner points (Harris and Stephens, 1988) and matched them by maximizing the normalized cross correlation (NCC) with subpixel refinement. In order to assess the SSE from the true inliers, as well as false positive (FP) and false negative (FN) counts, we manually classified the true inliers with the assistance of a helper script (see Fig. 7).

Fig. 7. Example close-up view of the script used for aiding a human classifier in inspecting the automatically found correspondences. Corresponding points are indicated by the red cross-hairs, and the motion vector field of nearby correspondences is also used to aid in spotting outliers. In many cases more careful analysis such as measuring or counting repeated patterns was also used.

In four out of the six image pairs (Table 1), SIMFIT had the lowest error. ASKC was lower for ANL and Kitchen pairs, but SIMFIT also did well here, and in a fraction of the time. For example, in the Kitchen pair, ASKC had squared error of 546.07 after 7.39 seconds, whereas SIMFIT had squared error of 677.18 after just 0.09 seconds.

Table 1. Experimental results of fundamental matrix estimation from real image correspondences. Runtime is measured in seconds. False positive (FP), false negative (FN), and SSE measures were computed based on the human classified inliers.

| Image Pair | | Runtime | SSE | FP | FN | Scale |
|---|---|---|---|---|---|---|
| Table (N=922, R=0.43)  | SIMFIT | **1.58** | **313.48** | 97 | 3 | 0.65 |
| | LMS+MSAC | 22.13 | 368.37 | 64 | 6 | 1.27 |
| | RANSAC EIS M | 20.02 | 448.66 | 22 | 74 | 0.31 |
| | ASSC | 10.15 | 368.37 | 36 | 24 | 0.71 |
| | ASKC | 11.92 | 368.37 | 32 | 37 | 0.71 |
| | STARSAC | 542.68 | 87190.9 | 285 | 35 | 9.27 |
| | RECON | 2.5 | 31875.3 | 84 | 75 | 4.67 |
| ANL (N=1181, R=0.25)  | SIMFIT | **0.84** | 1026.86 | 48 | 52 | 0.42 |
| | LMS+MSAC | 26.42 | 1786.2 | 56 | 62 | 0.79 |
| | RANSAC EIS M | 26.01 | 964.41 | 21 | 130 | 0.26 |
| | ASSC | 1.37 | 1786.2 | 61 | 38 | 1.11 |
| | ASKC | 15.54 | **925.43** | 34 | 84 | 0.70 |
| | STARSAC | 713.48 | 15217.9 | 185 | 37 | 3.55 |
| | RECON | 1.42 | 1588.02 | 49 | 40 | 0.92 |
| Kitchen (N=608, R=0.03)  | SIMFIT | **0.09** | 677.18 | 14 | 5 | 0.63 |
| | LMS+MSAC | 13.32 | 654.17 | 4 | 68 | 0.49 |
| | RANSAC EIS M | 12.81 | 615.63 | 2 | 135 | 0.26 |
| | ASSC | 0.58 | 552.56 | 5 | 33 | 0.66 |
| | ASKC | 7.39 | **546.07** | 4 | 80 | 0.53 |
| | STARSAC | 343.88 | 1015.31 | 4 | 273 | 0.20 |
| | RECON | 0.56 | 583.09 | 4 | 58 | 0.51 |
| Heinlein (N=453, R=0.05)  | SIMFIT | **0.12** | **274.45** | 0 | 8 | 0.43 |
| | LMS+MSAC | 9.92 | 2521.75 | 1 | 62 | 0.50 |
| | RANSAC EIS M | 9.54 | 2261.12 | 2 | 90 | 0.26 |
| | ASSC | 0.31 | 2521.75 | 2 | 30 | 0.87 |
| | ASKC | 5.52 | 371.565 | 0 | 43 | 0.62 |
| | STARSAC | 256.79 | 13011.5 | 14 | 67 | 2.19 |
| | RECON | 0.42 | 3640.52 | 0 | 32 | 1.27 |
| Desk (N=1149, R=0.08)  | SIMFIT | 0.63 | **1355.39** | 13 | 33 | 0.46 |
| | LMS+MSAC | 24.42 | 1492.02 | 7 | 95 | 0.62 |
| | RANSAC EIS M | 24.49 | 1400 | 3 | 187 | 0.27 |
| | ASSC | **0.58** | 1492.02 | 7 | 89 | 0.69 |
| | ASKC | 14.55 | 1454.49 | 6 | 107 | 0.69 |
| | STARSAC | 663.11 | 37411.9 | 93 | 0 | 15 |
| | RECON | 1.87 | 3278.3 | 16 | 107 | 1.22 |
| Livingroom (N=1854, R=0.05)  | SIMFIT | **0.32** | **587.20** | 26 | 6 | 0.41 |
| | LMS+MSAC | 39.27 | 1124.86 | 8 | 166 | 0.47 |
| | RANSAC EIS M | 38.84 | 666.51 | 3 | 265 | 0.22 |
| | ASSC | 0.73 | 1124.86 | 11 | 72 | 0.66 |
| | ASKC | 23.8143 | 710.847 | 9 | 136 | 0.49 |
| | STARSAC | 1073.86 | 76076.2 | 53 | 287 | 3.55 |
| | RECON | 5.27 | 1131.65 | 10 | 108 | 0.59 |

In general, one expects to find more inliers than actually exist when fitting the fundamental matrix because it is impossible to detect outlier correspondences that randomly happen to lie close to the epipolar line. Thus, our best indicator of algorithm performance is the objective SSE measure, which corresponds to negative log-likelihood.

SIMFIT had the fastest performance in 5 out of the 6 trials, taking significantly less than 1 second for all problems except Table, which had the highest outlier ratio, where it took 1.58 seconds. ASSC was often a close second in terms of performance, although it took 10 seconds on Table. The runtime performance of RECON was very reasonable due to the high inlier ratios, usually taking under 2 seconds, except for Livingroom where it took 5.27 seconds. STARSAC was by far the slowest algorithm, because it always performs a computationally intensive search through scale space to estimate model variances.

We do not have a ground truth reference for scale, but most algorithms estimate $\sigma$ in the range of 0.4 - 0.7 pixels for each image pair, which is almost certainly under-estimated. This is to be expected, because each correspondence has three degrees of freedom and four constraints, and thus there will always be a significant degree of over-fitting which allows the triangulated points to have lower error than the true points would. Thus, the distribution of fitting errors will be more peaked than predicted

by the $\chi^2$-distribution, leading to partially under-estimated scale by all algorithms. Nonetheless, the model estimates are still good.

## 7 Conclusions

RANSAC has proven to be an effective technique for overcoming large outlier ratios when a good threshold can be chosen, but it is sensitive to this choice. Several methods for augmenting RANSAC with automatic scale estimation have been previously proposed, but these methods tend to break down, or are too slow for many practical applications.

To overcome these limitations we have proposed the novel SIMFIT algorithm, which efficiently and reliably performs simultaneous scale estimation and model estimation without a breakdown point. SIMFIT is simple to implement, requires no new parameters (other than optional parameters for early termination), is reliable, and allows for adaptive sampling, keeping the runtime on-par with RANSAC for low as well as high outlier ratios.

Because SIMFIT is designed as a drop-in replacement for RANSAC, it can also be incorporated into other algorithms that use RANSAC as a subroutine. For example, the QDEGSAC (Frahm and Pollefeys, 2006) algorithm was designed to cope with quasi-degenerate data sets, and works by first running RANSAC to find a model that explains the data, and then estimating the codimension of the found model from the found inliers, and finally searching the remaining data for additional inliers that may provide the constraints necessary to make the model non-degenerate, if necessary. Thus, the initial RANSAC step can simply be replaced by SIMFIT to remove the need for *a priori* knowledge of scale.

Although we have demonstrated SIMFIT on some specific vision related problems, in addition to basic line fitting, we would like to stress that it is not specifically designed just for vision related tasks. We feel that the simplicity and generality of the method make it applicable to robust estimation in many other fields of science.

Finally, we note that there is room for future improvement in deriving a more advanced version of the $\chi^2$-distribution that accounts for increased peakedness due to over-fitting. Although the difference would be less than negligible for any normal model fitting problem (where errors are measured relative to a quantity that becomes increasingly over-determined from additional measurements), this would permit more accurate scale estimation for the special case of fundamental matrix estimation.

## References


P. J. Huber, Robust Statistics: A Review, Annals of Math and Statistics 43 (4) (1972) 1041–1067.

H. Wang, D. Suter, Robust adaptive-scale parametric model estimation for computer vision, Pattern Analysis and Machine Intelligence, IEEE Transactions on 26 (11) (2004a) 1459 – 1474, ISSN 0162-8828.

P. Rousseeuw, Robust Regression and Outlier Detection, Wiley, New York, 1987.

P. J. Rousseeuw, Least Median of Squares Regression, American Statistical Association 79 (388) (1984) 871–880.

P. Hough, Method and Means for Recognizing Complex Patterns, U.S. Patent 3.069.654, 1962.

R. O. Duda, P. E. Hart, Use of the Hough transformation to detect lines and curves in pictures, Commun. ACM 15 (1972) 11–15, ISSN 0001-0782.

L. Xu, E. Oja, P. Kultanen, A new curve detection method: randomized Hough transform (RHT), Pattern Recogn. Lett. 11 (1990) 331–338, ISSN 0167-8655.

M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (1981) 381–395.

P. Huber, Robust Statistics, John Wiley and Sons, Chichester, 1981.

P. H. S. Torr, A. Zisserman, MLESAC: a new robust estimator with application to estimating image geometry, Comput. Vis. Image Underst. 78 (1) (2000) 138–156.

O. Chum, J. Matas, J. Kittler, Locally Optimized RANSAC, in: DAGM-Symposium, 236–243, 2003.

O. Chum, T. Werner, J. Matas, Two-View Geometry Estimation Unaffected by a Dominant Plane, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, CVPR '05, IEEE Computer Society, Washington, DC, USA, 772–779, 2005.

J.-M. Frahm, M. Pollefeys, RANSAC for (Quasi-)Degenerate data (QDEGSAC), in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1, IEEE Computer Society, Washington, DC, USA, 453–460, 2006.

J. Matas, O. Chum, Randomized RANSAC with Td,d test, Image and Vision Computing 22 (10) (2004) 837 – 842, ISSN 0262-8856, british Machine Vision Computing 2002.

D. Capel, An effective bail-out test for RANSAC consensus scoring, Proc BMVC (2005) 629–638.

O. Chum, J. Matas, Optimal Randomized RANSAC, Pattern Analysis and Machine Intelligence, IEEE Transactions on 30 (8) (2008) 1472 –1482.

O. Chum, J. Matas, Matching with PROSAC - progressive sample consensus, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, 220 – 226 vol. 1, 2005.

D. Nister, Preemptive RANSAC for live structure and motion estimation, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, 199 –206 vol.1, 2003.

R. Raguram, J.-M. Frahm, M. Pollefeys, A Comparative Analysis of RANSAC Techniques Leading to Adaptive Real-Time Random Sample Consensus, in: Proceedings of the 10th European Conference on Computer Vision: Part II, Springer-Verlag, Berlin, Heidelberg, 500–513, 2008.

H. Wang, D. Suter, MDPE: A Very Robust Estimator for Model Fitting and Range Image Segmentation, IJCV 59 (2004b) 139–166.

P. H. S. Torr, D. W. Murray, The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix, Int. J. Comput. Vision 24 (1997) 271–300.

L. Fan, T. Pylvänäinen, Robust Scale Estimation from Ensemble Inlier Sets for Random Sample Consensus Methods, in: Computer Vision ECCV 2008, vol. 5304 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 182–195, 2008.

H. Chen, P. Meer, Robust regression with projection based M-estimators, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, vol. 2, 878–885, 2003.

S. Rozenfeld, I. Shimshoni, The Modified pbM-Estimator Method and a Runtime Analysis Technique for the RANSAC Family, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 17, 1113–1120, 2005.

R. Subbarao, P. Meer, Heteroscedastic Projection Based M-Estimators, in: CVPR Workshops, IEEE Computer Society Conference on, 2005.

R. Subbarao, P. Meer, Beyond RANSAC: User Independent Robust Regression, in: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, CVPRW '06, IEEE Computer Society, Washington, DC, USA, ISBN 0-7695-2646-2, 2006.

H. Wang, D. Mirota, G. D. Hager, A Generalized Kernel Consensus-Based Robust Estimator, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (1) (2010) 178–184.

J. Choi, G. G. Medioni, StaRSaC: Stable random sample consensus for parameter estimation, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, IEEE, ISBN 978-1-4244-3992-8, 675–682, 2009.

R. Raguram, J.-M. Frahm, RECON: Scale-Adaptive Robust Estimation via residual consensus, in: Proc. of the 2011 Intl. Conf. on Computer Vision, 1299–1306, 2011.

R. Toldo, A. Fusiello, Automatic Estimation of the Inlier Threshold in Robust Multiple Structures Fitting, in: Image Analysis and Processing - ICIAP 2009, 123–131, 2009.

T.-J. Chin, H. Wang, D. Suter, Robust fitting of multiple structures: The statistical learning approach, in: Computer Vision, 2009 IEEE 12th International Conference on, ISSN 1550-5499, 413 –420, 2009.

H. Wang, T.-J. Chin, D. Suter, Simultaneously Fitting and Segmenting Multiple-Structure Data with Outliers, Pattern Analysis and Machine Intelligence, IEEE Transactions on 34 (6) (2012) 1177 –1192.

N. Snavely, S. M. Seitz, R. Szeliski, Photo tourism: Exploring photo collections in 3D, in: SIGGRAPH Conference Proceedings, ACM Press, New York, NY, USA, 835–846, 2006.

N. Snavely, S. M. Seitz, R. Szeliski, Modeling the World from Internet Photo Collections, Int. J. Comput. Vision 80 (2008) 189–210.

K. Kanatani, Statistical Optimization for Geometric Computation: Theory and Practice, Elsevier Science Inc., New York, NY, USA, ISBN 0444824278, 1996.

D. D. Dyer, Estimation of the Scale Parameter of the Chi Distribution Based on Sample Quantiles, Technometrics 15 (3) (1973) 489–496.

R. I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, ISBN: 0521540518, second edn., 2004.

X. Yu, T. Bui, A. Krzyzak, Robust Estimation for Range Image Segmentation and Reconstruction, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1994) 530–538.

K.-M. Lee, P. Meer, R.-H. Park, Robust Adaptive Segmentation of Range Images, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 200–205, ISSN 0162-8828.

A. Bab-Hadiashar, D. Suter, Robust segmentation of visual data using ranked unbiased scale estimate, Robotica 17 (1999) 649–660.

J. Branham, R. L., Alternatives to least squares, Astronomical Journal 87 (1982) 928–937.

D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (5) (2002) 603 –619.

R.C.A. Victor Company, Dallin Aerial Survey Company Photographs, HagleyID 70.200.08432, Hagley Museum and Library, 1935a.

R.C.A. Victor Company, Dallin Aerial Survey Company Photographs, HagleyID 70.200.08434, Hagley Museum and Library, 1935b.

B. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1986.

J. P. Nolan, Stable Distributions - Models for Heavy Tailed Data, Birkhauser, Boston, in progress, Chapter 1 online at academic2.american.edu/∼jpnolan, 2011.

C. Harris, M. Stephens, A Combined Corner and Edge Detector, in: Proceedings of the 4th Alvey Vision Conference, 147–151, 1988.