RECON: Scale-Adaptive Robust Estimation via Residual Consensus

Rahul Raguram　　Jan-Michael Frahm
Department of Computer Science
University of North Carolina at Chapel Hill
{rraguram, jmf}@cs.unc.edu

Abstract

In this paper, we present a novel, threshold-free robust estimation framework capable of efficiently fitting models to contaminated data. While RANSAC and its many variants have emerged as popular tools for robust estimation, their performance is largely dependent on the availability of a reasonable prior estimate of the inlier threshold. In this work, we aim to remove this threshold dependency. We build on the observation that models generated from uncontaminated minimal subsets are "consistent" in terms of the behavior of their residuals, while contaminated models exhibit uncorrelated behavior. By leveraging this observation, we then develop a very simple, yet effective algorithm that does not require apriori knowledge of either the scale of the noise, or the fraction of uncontaminated points. The resulting estimator, RECON (REsidual CONsensus), is capable of elegantly adapting to the contamination level of the data, and shows excellent performance even at low inlier ratios and high noise levels. We demonstrate the efficiency of our framework on a variety of challenging estimation problems.

1. Introduction

A ubiquitous task in computer vision is that of estimating the parameters of a model from data that may be contaminated with measurement noise and outliers. Random Sample Consensus (RANSAC) [12] is one of the most widely used techniques for problems of this form. The basic algorithm operates in a hypothesize-and-verify loop: minimal subsets of points are repeatedly sampled from the data and used to generate model hypotheses. Each model is subsequently scored based on its support, given by the number of points whose error with respect to the model is less than some fixed threshold. The goal in RANSAC is to return the model that maximizes the cardinality of this support.

In addition to its remarkable simplicity, one compelling reason for RANSAC's widespread adoption is its ability to tolerate large levels of data contamination. It is worth noting, however, that the robustness of the algorithm is contingent on an appropriate choice of inlier threshold (or, equivalently, *a priori* knowledge of the variance of measurement noise). Since models are scored based on their support, it is precisely this parameter that allows RANSAC to distinguish between good and bad model fits. While it may indeed be possible to often "guess" a reasonable threshold value, there are applications (such as the alignment of 3D reconstuctions with arbitrary scale), where this choice is much less obvious. In addition, it has been shown that if the chosen threshold deviates significantly from the true value, RANSAC can provide biased, or simply incorrect results [5].

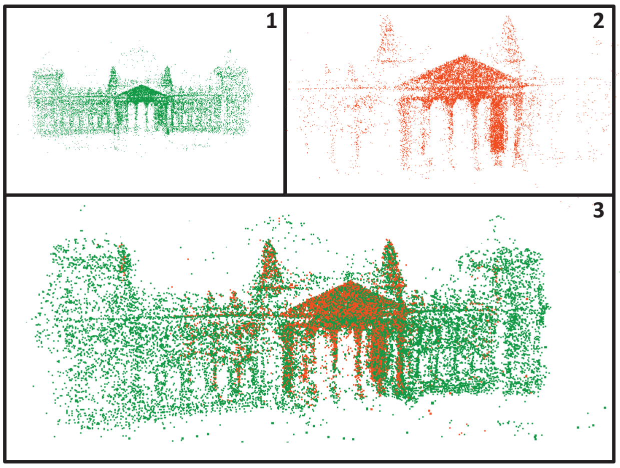


Figure 1: [1-2] Two sparse 3D submodels obtained from a structure-from-motion system. The two reconstructions are to be merged into a common reference frame via a 3D similarity transformation. Since the scale of each reconstruction is arbitrary, a good choice of threshold is not obvious. [3] Merged reconstruction (best viewed in colour) obtained using our method, requiring no prior knowledge of noise scale or inlier ratio.

In this work, we present a novel robust estimation framework that is agnostic to the threshold parameter. Instead, the algorithm we present is based on a simple observation: that the residual errors for "good" models are in some way consistent with each other. We show that it is possible to efficiently identify this stable behaviour by exploiting residual ordering information coupled with simple non-parametric statistical tests. This leads to an very simple, yet effective, algorithm whose performance over a range of noise levels and inlier ratios is comparable to that of RANSAC, whilst

978-1-4577-1102-2/11/$26.00 ©2011 IEEE

completely eliminating the inlier threshold parameter.

To summarize the main contributions of this work: (1) we develop a new framework for robust estimation, based on the behaviour of residuals. (2) The method does not require an inlier threshold, or a worst-case assumption about the level of data contamination. This is a strongly distinguishing characteristic compared to current techniques, which require at least one of these assumptions to be made. (3) The effectiveness of the framework is demonstrated on a variety of challenging estimation problems covering a range of inlier ratios, noise levels, and model complexities.

2. Related Work

Since the introduction of the original RANSAC algorithm, a number of popular variants have emerged [9, 20, 7, 11, 13, 8, 21, 22, 19, 24]. Orthogonal to these advances, there have also been attempts to develop robust estimators capable of operating without a threshold parameter.

Since scoring models based on their support requires a threshold, one possible approach is to to optimize a different cost function. Least Median of Squares (LMedS) [23], scores models based on their median error residual, returning the model with lowest score. However, this breaks down when the data consists of more than 50% outliers, since the median residual is no longer informative. MINPRAN [26] searches for a combination of model parameters and inliers that are least likely to have occured by chance; however, this assumes knowledge of the dynamic range of the outlier data. MUSE [18] and ALKS [15] are related techniques that minimize the k^{th}order statistics of the squared residuals. It has been noted [28] that their ability to handle large outlier ratios is limited. ASSC [28] uses a mean-shift algorithm to detect modes in the residual distribution, using this to estimate the noise scale. However, this can be computationally expensive, particularly as the model complexity increases [29]. It is worth noting that while many of the above techniques have been evaluated for segmenting multiple structures from range data, it has been observed [27] that extending them to other estimation problems, such as multi-view geometric relations, can be challenging. The recently proposed StaRSaC algorithm [5] relaxes the requirement of a fixed threshold, by exhaustively testing all choices in some range. For each of T thresholds, RANSAC is executed K times, and a value is chosen that shows minimum "variance of parameters". Even for reasonable values of T and K, this quickly becomes computationally infeasible (e.g., testing $T = 30$ thresholds with K = 30 requires almost three orders of magnitude more computation than RANSAC).

The two recent approaches that are most similar in spirit to ours are those of Zhang et al. [29] and Kernel-Fitting [3]. Both these works use the observation that for a given point, the distribution of residuals with respect to a large set of randomly generated models can be used to reveal whether the point is an outlier or an inlier. In particular, the kernel-based measure introduced in [3] has been shown to yield good results for the problem of fitting multiple structures to noisy data [2, 4]. However, one limitation of these strategies is their implicit reliance on sampling a "sufficient" number of uncontaminated models. It is worth noting that models that are contaminated carry no reliable information about the validity of a data point not contained in the minimal sample itself. This is intuitive: the error of any inlier data point to a bad model is arbitrary. Thus, it becomes imperative to sample enough models for the distribution of residuals to be sufficiently "informative". We take a different view: instead of the residual distributions of *points*, we observe the residual distributions of *models*. The intuition here is that two uncontaminated models will be "alike" in some way, while contaminated models will disagree in an unstructured way. This results in a simpler, yet much more efficient algorithm.

Finally, we note that one fundamental advantage of our technique compared to the methods discussed in this section [23, 26, 18, 15, 28, 29, 3, 2, 4] is the ability to adaptively stop evaluation. In standard RANSAC, the current best inlier count can be used to adaptively set the maximum number of trials. However, in the absence of a threshold, current techniques either require knowledge of the inlier ratio, or must run the algorithm under a worst-case assumption–in other words, drawing enough samples to guarantee success for the *worst* inlier ratio that could potentially occur in the data. RECON does not have this limitation; we are able to adapt "on-the-fly" to the level of data contamination, without requiring any data-specific parameter tuning.

3. The Approach

3.1. Residual Consensus

The key idea behind RANSAC is that a model generated from an uncontaminated minimal sample (i.e., containing only inliers) will have larger support than a contaminated model (generated from samples containing at least one outlier). Provided enough random samples have been drawn to guarantee, with some confidence, that at least one of the minimal samples is free from outliers, RANSAC will terminate successfully, returning the corresponding model with highest support. Observe that RANSAC uses the threshold t to essentially distinguish between good and bad model estimates. In the absence of a threshold, however, it becomes harder to make this distinction. Our technique is based on the intuition that, near the true threshold value, uncontaminated models capture a stable set of points (i.e., the set of inliers). On the other hand, contaminated models do not demonstrate this consistent behaviour. In other words, looking at the relationship between *subsets* of models can reveal useful information about the validity of the models themselves. To formalize this intuition, we now inspect the behaviour of uncontaminated vs. contaminated models.

3.1.1 Uncontaminated models: Given an uncontaminated model and access to the true noise variance, it is *always* possible to compute a region containing a fixed fraction α of the total number of inliers [14, 22]. To see why this is the case, consider the typical assumption that the measurement error for inliers is Gaussian, with zero mean and variance σ^2. Let our estimate of the (*a priori* unknown) variance be $\hat{\sigma}^2$. Under the Gaussian assumption, the squared point-model errors between an inlier and the "true" model can be represented as a sum of m squared Gaussian variables, where m is the codimension of the model – i.e., a chi-square distribution with m degrees of freedom. Thus, in order to recover a fraction α of inliers, an appropriate threshold t can be computed as $t^2 = \chi_m^{-1}(\alpha)\sigma^2$, where $\chi_m^{-1}(\cdot)$ is the inverse cumulative chi-square distribution. In other words, if our estimated $\hat{\sigma}^2$ is close to the true σ^2, then using the threshold t for an uncontaminated model will result in a fraction α (usually set to 0.95 or 0.99) of the inliers being found. For the case of line fitting, this is shown in Fig. 2(a).

Note, however, that to be strictly correct, one must take the uncertainty of the estimated models into account. Since models are generated from noisy data points, this error propagates to the model itself [9, 22]. As a consequence, when using the "ideal" threshold, uncontaminated models do not always capture a fixed fraction α of inliers. However, if we characterize the uncertainty of a model $\mathbf{M}$ via its covariance matrix $\Sigma_{\mathbf{M}}$ and explicitly perform error propagation, this can be remedied [22]. For instance, when fitting a line, we can derive an expression for the covariance matrix Σ_L of the line [16]. To capture a fraction α of true inliers, we now need to consider the α-percentile error envelope defined by Σ_L, which is in the form of a hyperbola (Fig. 2(b)).

From the discussion above, given that a fraction α of the inliers can always be recovered for an uncontaminated model, now consider *a pair* of uncontaminated models. It can be seen that the expected fraction of total inliers in the intersection of the two error envelopes is given by α^2. For typical values of α close to 1.0, note that this expected fraction is close to α, implying that uncontaminated models capture a stable set of points with respect to each other (Fig. 2(c)). In other words, if $\hat{\sigma}^2 \approx \sigma^2$, uncontaminated models will have a fraction $\approx \alpha$ of their inliers in common. We refer to these models as being *α-consistent*[1].

3.1.2 Contaminated models: Assuming unstructured outliers, observe that models generated from contaminated subsets have arbitrary parameters. Given our estimate of the noise variance $\hat{\sigma}^2$, we can break down the behaviour of contaminated models into two scenarios, based on the relationship between the estimated $\hat{\sigma}^2$, and the true σ^2.

(1) $\hat{\sigma}^2 \leq \sigma^2$: First, assume $\hat{\sigma}^2 = \sigma^2$. As before, we can use the true noise variance to compute α-percentile envelopes, but they now capture essentially arbitrary data points (Fig. 2(d)). In particular, the probability of a pair of contaminated models being α-consistent (i.e., having significant overlap in their inlier sets) tends to zero, since this would amount to picking essentially the same minimal sample twice. To see this, assume that we have N data points, with I inliers. The probability of picking an uncontaminated minimal sample of size s is given by

$$P_{good} = \prod_{i=0}^{s-1} \frac{I-i}{N-i} \approx (I/N)^s \tag{1}$$

Now, assume we have generated a contaminated model. Picking a second model that happens by chance to be consistent with the first is governed by the probability

$$P_{bad} = \prod_{i=0}^{s-1} \frac{I'-i}{N-i} \tag{2}$$

where I' is the number of points supporting the first "bad" model. Assuming unstructured outliers, we have $I' \ll I$, implying that $P_{bad} \ll P_{good}$. Note that exactly this same reasoning holds for the case of $\hat{\sigma}^2 < \sigma^2$.

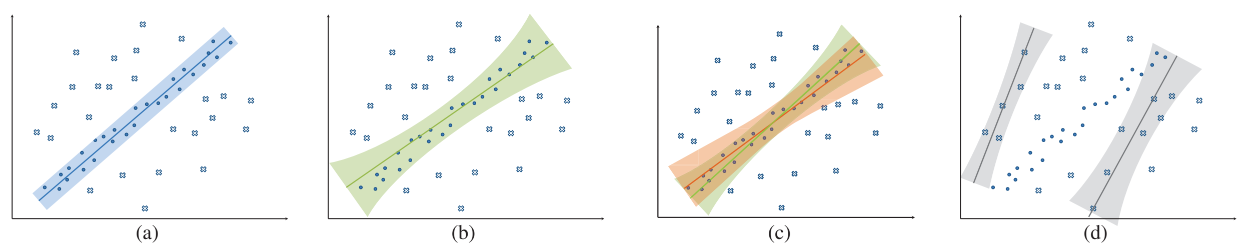
(2) $\hat{\sigma}^2 > \sigma^2$: The extent of the α-percentile envelopes is governed by $\hat{\sigma}^2$. In particular, for $\hat{\sigma}^2 \gg \sigma^2$, the computed error envelopes might become so large that they include all data points. This leads trivially to α-consistency for any pair of models, uncontaminated or otherwise, since all points fall in the intersection. If no prior is available on the potential values of $\hat{\sigma}$, we must use additional constraints to determine whether we have an overestimate of σ. We do so by leveraging knowledge of the distribution of residuals.

Given two models $\mathbf{M}_1$ and $\mathbf{M}_2$ that are found to be α-consistent, we know that there is significant overlap in their inlier sets. Assuming that the size of this overlap is n, we have two residual vectors $R^1 = \{r_1^1, r_2^1, ..., r_n^1\}$ and $R^2 = \{r_1^2, r_2^2, ..., r_n^2\}$, measured from each model to the n points. Under typical assumptions, we know that the distribution of inlier residuals from an uncontaminated model approximates a chi-square distribution. In other words, if $\mathbf{M}_1$ and $\mathbf{M}_2$ are uncontaminated and $\hat{\sigma} \approx \sigma$, the vectors R^1 and R^2 can be thought of as two independent random samples drawn from a *common* underlying chi-square distribution. On the other hand, when one or more of the models are contaminated and $\hat{\sigma} \gg \sigma$, the intersection will contain an *arbitrary combination* of inliers and outliers. Thus, a statistical test that evaluates the similarity between distributions can reveal useful information about the nature of the models. The Kolmogorov-Smirnov (KS) [1] and Anderson-Darling (AD) [25] tests are statistical tools that may be used to compare a sample with a reference distribution or to compare two empirical samples. The KS test computes the statistic

$$D_{n,n} = \sqrt{n/2} \sup_x |F_{1,n}(x) - F_{2,n}(x)|, \tag{3}$$

where n is the number of points, $F_{1,n}(x)$ is an empirical distribution function and $F_{2,n}(x)$ is either a reference distribution (one-sample KS test) or a second empirical distribution function (two-sample KS test). We wish to test the

[1] For simplicity, discussions of α-consistency will assume $\alpha = 0.99$.



(a) (b) (c) (d)

Figure 2: (a) Error bound that captures $\alpha = 99\%$ of true inliers with respect to a noise-free, uncontaminated model. Using the inverse chi-square cdf, this gives $t = 2.57\sigma$. (b) Error envelope, following model covariance estimation, capturing $\alpha = 99\%$ of the inliers for a model generated from noisy inliers. (c) Overlap between two uncontaminated models, with a fraction $\approx \alpha$ of their inliers in common (d) Error envelopes for contaminated models, capturing random data points. (Figure best viewed in colour)

null hypothesis H_0 that the two samples R^1 and R^2 are from the same chi-square distribution. H_0 is rejected at a level β (typically 1% or 5%) if $D_{n,n} > C_\beta$, where C_β is the critical value of the Kolmogorov distribution (for details, refer [1]). In other words, if the residuals offer strong evidence against the null hypothesis, the models can be discarded as invalid.

Summarizing the entire discussion thus far:

1. Given the true σ, the probability of contaminated models being α-consistent is vanishingly small.
2. Conversely, at the true σ, a pair of uncontaminated models will be α-consistent.
3. If a pair of models is found to be α-consistent for some σ, their inlier residuals must agree on a common underlying chi-square distribution.

The above observations are important because they direct us towards an algorithm for fitting models *in the absence of true scale*. In brief: iterate over possible values of σ, and pick the lowest σ that gives rise to a set of α-consistent models whose residuals are statistically likely to correspond to the same distribution. The algorithm is outlined in Alg 1. In practice, note that we search for a set of K ($= 3$ in our experiments) mutually consistent models to provide robustness to some degree of structure in the outliers. We have implemented and tested the algorithm sketched above. While our experiments verify the correctness of the approach, we note that it has the drawback of being computationally expensive. The algorithm requires us to examine all possible σ values and rigorously perform covariance estimation for all models. In the context of these limitations, we now present a practical algorithm that efficiently exploits *residual ordering* information to achieve the same results.

Algorithm 1 Iterative robust estimation via α-consistency

```
Repeat 1-3 until $K$ mutually $\alpha$-consistent models found
1. Hypothesis generation:
Hypothesize model $\mathbf{M}_i$ from random minimal sample
2. Model verification:
Compute and store the residuals of $\mathbf{M}_i$ to all data points
3. Test $\alpha$-consistency
for all $\hat{\sigma}$ in range $(\sigma_{min}...\sigma_{max})$ do
    for all prior models $\mathbf{M}_j$ in $\{\mathbf{M}_1, \mathbf{M}_2..., \mathbf{M}_{i-1}\}$ do
        Use $\hat{\sigma}$ to compute error envelopes for $\mathbf{M}_i$ and $\mathbf{M}_j$
    if set of $K$ pairwise $\alpha$-consistent models found then
        if inlier residuals pass the KS statistical test then
            Return $\hat{\sigma}$, models and set of inliers
```

3.2. Efficient Residual Consensus

Given a model $\mathbf{M}_i$, consider its residuals to all N data points, $R^i = \{r_1^i, r_2^i, ..., r_N^i\}$. Assume that for each model, we sort the residuals to be in increasing order, and retain the indices of the sorted points. In other words, we have a set $S^i = \{\lambda_1^i, \lambda_2^i, ..., \lambda_N^i\}$, representing a permutation of the indices $1, 2, ..., N$ such that $r^i_{\lambda^i_1} \leq r^i_{\lambda^i_2} \leq ... \leq r^i_{\lambda^i_N}$. We define the partial overlap set $\Theta_n^{i,j}$ as

$$\Theta_n^{i,j} = S_{1:n}^i \bigcap S_{1:n}^j \tag{4}$$

where i and j denote two models $\mathbf{M}_i$ and $\mathbf{M}_j$ and $S_{1:n}^i$ is the subset containing the first n points of S^i. We also define the *normalized partial overlap*, $\theta_n^{i,j}$, to be the scalar quantity

$$\theta_n^{i,j} = \frac{1}{n}|\Theta_n^{i,j}| \tag{5}$$

This represents the number of common elements when considering the first n points, normalized by the subset size. We finally define the expected fraction of common points when considering a subset of size n, E_{θ_n},

$$E_{\theta_n} = E[\theta_n^{i,j}] = \frac{1}{n}\sum_{k=0}^{n} kp(|\Theta_n^{i,j}| = k) \tag{6}$$

One way to think about the index sets $S_{1:n}^i$ is the following. By choosing a subset of size n, we are implicitly defining a threshold t_n^i (and thus, some $\hat{\sigma}$), and then selecting from the vector R^i, those points with $r^i < t_n^i$. The approach we take is based on the intuition that for uncontaminated models, inlier indices should loosely group together at the front of set S^i. On the other hand, contaminated models should represent random permutations of the point indices. We shall now look more closely at this behaviour.

1. Uncontaminated models: Given sorted index sets S^i and S^j for a pair of uncontaminated models, it is evident that inlier indices should appear towards the front of these sets (by definition, inliers are closer to uncontaminated models than outliers). Note that due to model noise, the *relative ordering* of the inlier indices is arbitrary. However, as we consider progressively larger subsets $S^i_{1:n}$ and $S^j_{1:n}$, this defines implicit thresholds t^i_n and t^j_n . From Section 3.1, we know that when both t^i_n and t^j_n are near the true threshold, the subsets $S^i_{1:n}$ and $S^j_{1:n}$ each contain a significant fraction α of the inliers. In other words, the normalized overlap $\theta^{i,j}_n$ should approach α^2 near the true threshold.

2. Contaminated models: Given a pair of contaminated models with arbitrary parameters, we do not expect to see consistent structure when inspecting their residuals. In other words, after sorting by residuals, the sets S^i and S^j for a pair of bad models are likely to be random permutations of the point indices $1, 2, ..., N$. The probability that the subset $S^i_{1:n}$ for model $\mathbf{M}_i$, has k elements *by chance* in common with the subset $S^j_{1:n}$ can be derived as:

$$p(|\Theta^{i,j}_n| = k) = \binom{n}{k}\binom{N-n}{n-k} \Big/ \binom{N}{n} \qquad (7)$$

Substituting into Eqn (6), this results in

$$E_{\theta_n} = n/N \qquad (8)$$

In other words, the normalized overlap between two contaminated models approaches α^2 only as $n \approx N$. This poses a question: how do we distinguish this case from that of two good models where the inlier ratio is close to 100%? Some consideration will reveal that this ambiguity corresponds exactly to the discussion in Sec. 3.1.2, of the case $\hat{\sigma} \gg \sigma$ where the error envelopes include essentially all data points. Note that given a set of inlier residuals $\{r^i_1, r^i_2, ..., r^i_n\}$, a robust estimate of the noise variance may be easily computed [23, 18, 27] as

$$\hat{\sigma} = C\sqrt{\text{median}\{{r^i_m}^2\}}, \text{where}\, m = 1, ..., n \qquad (9)$$

and C is a constant that makes $\hat{\sigma}$ unbiased for a particular target distribution of residuals. For instance, $C = 1.4286(1 + 5/(n - s))$ for Gaussian distributed residuals, where s is the minimal sample size. When a pair of bad models is found to be α-consistent, the corresponding scale estimate from Eqn. (9) is typically dramatically larger than the true value. Thus, if a loose estimate (even up to an order of magnitude) of the maximum possible scale is available, it can be used to quickly discard bad model pairs. Alternatively, the same KS statistical test from Section 3.1 can be used to determine the validity of the residuals. Note that all the above arguments hold when one of the two models is bad; the probability of a random model producing the same sorted indices as a good model is also given by Eqn. (7).

To empirically verify our analysis of the behaviour of contaminated vs. uncontaminated models, we conducted the following experiments. Synthetic data was generated for three model-fitting examples: line, homography and 3D similarity estimation. The inlier ratio in all cases was fixed at 40% and $N = 1000$ points were generated. We then plotted the normalized partial overlap $\theta^{i,j}_n$, between random pairs of good models and bad models (Fig. 3). The following observations are pertinent:

1. For pairs of uncontaminated models (top row in Fig. 3), we observe that when $n \approx I$ (close to the true inlier ratio of 40% in these experiments), the value of $\theta^{i,j}_n$ approaches α^2. This supports our observation regarding α-consistency scores for uncontaminated models.

2. For pairs of contaminated models (bottom row in Fig. 3), it can be seen that there is close agreement between the theoretical and obtained plots. In particular, $\theta^{i,j}_n \approx \alpha^2$ only for $n \approx N$, as predicted in Eqn (8). This validates the argument that the residuals for contaminated models are unstructured.

3.3. RECON: Algorithm Description

The details of the Residual Consensus (RECON) algorithm are presented in Algorithm 2. Some notes follow:

Algorithm 2 RECON: REsidual CONsensus

```
Repeat 1-3 until $K$ consistent models found
1. Hypothesis generation:
Hypothesize model $\mathbf{M}_i$ from random minimal sample
2. Model verification:
Sort residuals, store sorted index set $S^i$ and residual set
3. Test $\alpha$-consistency
for all prior models $\mathbf{M}_j$ in $\{\mathbf{M}_1, \mathbf{M}_2 ..., \mathbf{M}_{i-1}\}$ do
  if $\mathbf{M}_i, \mathbf{M}_j$ are $\alpha$-consistent ($\theta^{i,j}_n \geq \alpha^2$ for some $n$) then
    if $\hat{\sigma} < \sigma_{max}$ (Eqn. 9) or KS test passed (Eqn. 3) then
      Add to set of $\alpha$-consistent models
      if $K$ mutually consistent models found then
        Return set of $K$ models and inliers
4. Refine to obtain final solution
Sort inliers by mean of their residuals to the $K$ models
Generate $M$ non-minimal models using PROSAC-based
sampling of the sorted inlier set
Recompute pairwise $\theta^{i,j}_n$, select minimum $n$ that satisfies
$\theta^{i,j}_n \geq \alpha^2$ for $i, j \in 1, ..., M$.
Re-estimate final model using the best inlier set.
```

1. As in Alg. 1, we look for K pairwise α-consistent models to provide robustness to structure in the outliers. In all our experiments, setting $K = 3$ was sufficient in practice.

2. From a computational viewpoint, once enough samples have been drawn to rule out high inlier ratios (say, $\varepsilon > 90\%$), the residual test can be bypassed. This is because a bad model pair has $\theta^{i,j}_n \approx \alpha^2$ only for $n \approx N$ (Eqn. 8), which implies that the resulting inlier ratio is close to 100%. If enough models are drawn to rule out this case, α-consistent models with large n can be directly excluded.

3. For clarity, we have thus far deferred the issue that a small number of outliers may satisfy our (implicit) threshold constraints, and be included in the overlap sets. In addition, the index sets $S_{1:n}$ can be viewed as an approximation to

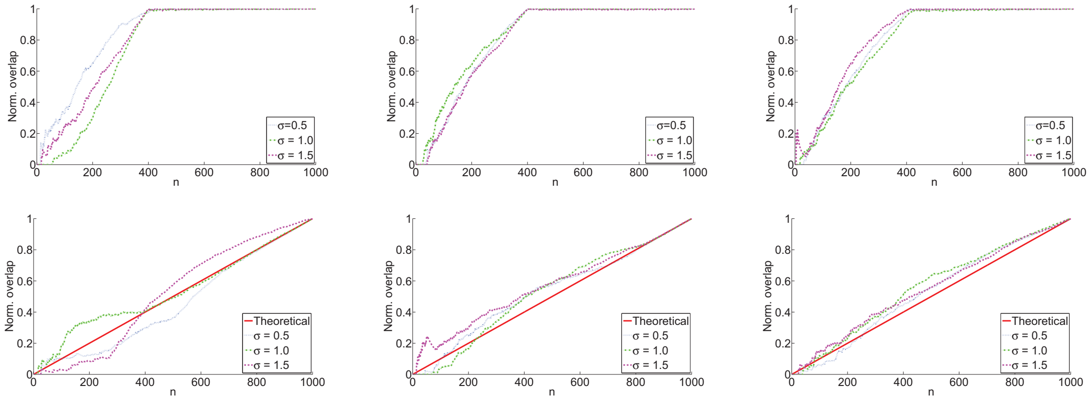


Figure 3: (Left-Middle-Right) 2D-line, homography, 3D similarity transform estimation. (Top row) Normalized overlap $\theta_n^{i,j}$, vs. subset size n for pairs of uncontaminated models. Note that close to the true inlier ratio (40%), $\theta_n^{i,j} \approx \alpha^2$. . (Bottom Row) Normalized overlap $\theta_n^{i,j}$ vs. n for random pairs of contaminated models. Also overlaid is the predicted value from Eqn. (8). Note that $\theta_n^{i,j} \approx \alpha^2$ as $n \approx N$.

the covariance envelopes in Sec 3.1. The practical implication of these two points is that we sometimes overestimate the inlier set. In other words, when $\theta_n^{i,j} \approx \alpha^2$, we have $n > I$, implying that we have included additional outliers. Observe that the source of this problem is noise in the minimal sample. To mitigate this, we generate M *non-minimal* samples ($M = 20$ in our experiments) from the returned inlier set, using PROSAC-based non-uniform sampling [7] (Step 4 in Alg. 2). Reestimating the overlap should now result in $n \approx I$, since we are explicitly compensating for model noise. As will be shown in Sec. 4, this results in the final set of inliers being very close to the ground truth.

4. An additional speedup is obtained by discarding models that violate certain basic checks (such as oriented constraints, for geometry estimation). As noted in [10], this can result in upwards of 70-80% of the models being discarded. This yields a much smaller core set of models that needs to be checked (Step 3 in Alg. 2).

4. Experiments

In this section, we evaluate RECON's performance for a number of estimation problems. We stress that RECON does not require an inlier threshold, or any model-specific parameter tuning. The only required parameters are α and K, which are set to 0.99 and 3 respectively, for all experiments. Finally, note that RECON is able to adapt to the level of data contamination, and does not require the generation of an *a priori* fixed number of hypotheses.

4.1. Synthetic data:

We exhaustively evaluated the performance of RECON on four estimation problems: 2D line, 3D plane, homography (**H**), and fundamental matrix (**F**). The inlier ratio ε was varied over a wide range [0.25-1.0], in steps of 0.05. For the line and plane experiments, we generated points within a square (resp., cube) of side 500, and added zero-mean Gaussian noise to the inliers, with σ varying from 0.5-6.0. For the **H** and **F** tests, we generated synthetic 2D correspondences, varying σ from 0.5-4.0 (roughly the accuracy of current feature detectors [17]). For each (ε, σ) pair, the total number of points was randomly chosen in the range [500, 2000].

Given the large number of parameter combinations tested, we present a representative selection of the results, noting briefly that RECON provided accurate results over the entire test dataset. Table 1 provides a baseline comparison of RECON with RANSAC and LMedS, listing the number of samples drawn (k) and the mean error (err), averaged over 500 runs. RANSAC$_F$ and LMedS$_F$ denote "realistic" cases, where the algorithms are run using commonly chosen parameter values. We use $\sigma = 1.0$ to set the RANSAC$_F$ threshold, and simply use the median score in LMedS$_F$, with the number of trials computed using a 50% inlier ratio. To denote the "ideal" case, we also run the algorithms with *true* parameter values which, of course, are typically unknown. For these cases, denoted by RANSAC$_T$ and LMedS$_T$, the RANSAC threshold is set using the true σ, and LMedS is adapted to switch between the median or the 25% lower quartile, depending on the true ε.

Perhaps the most important observation to be made from the results is that RECON is able to provide results with virtually the same accuracy as RANSAC$_T$ and LMedS$_T$, while requiring *no* parameter tuning. In addition, when RANSAC and LMedS are run with incorrect parameter values, their performance deteriorates significantly. An interesting statistic to note from Table 1 is that the number of trials for RECON is appreciably lower than for RANSAC, even when RANSAC is run using the true threshold value. Indeed, as shown in Fig. 4(a), the number of samples drawn by RECON is significantly close to that predicted by the standard RANSAC stopping criterion using a 95% confidence level. This is not a coincidence: it can be shown [6] that for a 95% confidence in the solution, three uncontaminated samples are drawn on average in RANSAC before the confidence in the solution is achieved. Since we look for $K = 3$ consistent models, it is precisely these samples that RECON finds. RANSAC has the opposite problem: since noisy models have lower support, this means more samples must be drawn to meet the stopping criterion [9].

Fig. 4(b) plots the fraction of true inliers found by RECON. Note that in all cases, RECON returns close to 100% of the true inliers, while RANSAC suffers, particularly for higher noise levels. This relates to the discussion in Sec. 3.1.1. Using a fixed threshold with noisy models implies that not all inliers will be found. On the other hand, since we look for *consistent* sets of inliers, and do not impose a fixed threshold, a higher fraction of true inliers is returned.

4.2. Real data

We also tested RECON's performance on real data, for four model fitting problems: homography, essential matrix, fundamental matrix and 3D similarity. Since the ground truth noise variance is unknown, we use the commonly used setting of $\sigma = 1.0$ for the two-view estimation problems. Setting the threshold for the 3D similarity estimation problem is much more difficult, since the submodels are reconstructed up to an arbitrary scale. For this case, the RANSAC threshold was determined using an exhaustive trial-and-error approach. The results are presented in Table 2. Note that the performance of RECON is comparable to RANSAC, while requiring *no threshold*. Our implementation of RECON, though as yet unoptimized for speed, has runtimes between 30 milliseconds and 2.5 seconds for the experiments in Table 2, while the corresponding range for RANSAC is 22ms–1.7s. RECON's ability to adaptively stop evaluation is especially favourable when contrasted with current techniques that use either a pessimistic estimate of the number of hypotheses (e.g., 10000 in [28]) or a fixed time window (e.g., 60s in [4]).

5. Conclusion

In this paper, we have developed a new, threshold-free framework for robust estimation. In contrast to prior work, we use *consistency* of models as a cue to detect model estimates that are close to the true solution. We develop the theoretical framework, and then present a practical algorithm, RECON, that can be used to efficiently perform robust estimation. The algorithm is very simple, easy to implement, and effective in practice. RECON produces results of the same quality as algorithms that specifically use fine-tuned parameter estimates. An additional benefit of the flexible consensus measure is the ability to elegantly adapt the number of samples drawn to the contamination level of the data. In future work, we plan to address the effect of structured outliers on the estimation results.

Acknowledgements: The authors thank Christopher Zach for providing valuable suggestions. This research was supported by NSF grant IIS-0916829, DOE grant DEFG52-08NA28778 and Navy SPAWAR.

References

[1] I. M. Chakravarti, R. G. Laha, and J. Roy. *Handbook of Methods of Applied Statistics*, volume I. 1967.

[2] T. J. Chin et al. The Ordered Residual Kernel for Robust Motion Subspace Clustering. In *NIPS*, 2009.

[3] T.-J. Chin, H. Wang, and D. Suter. Robust fitting of multiple structures: The statistical learning approach. In *ICCV*, 2009.

[4] T.-J. Chin, J. Yu, and D. Suter. Accelerated hypothesis generation for multi-structure robust fitting. In *ECCV*, 2010.

[5] J. Choi and G. Medioni. StaRSaC: Stable random sample consensus for parameter estimation. *CVPR*, 2009.

[6] O. Chum. *Two-View Geometry Estimation by Random Sample and Consensus*. PhD thesis, CTU, Prague, 2005.

[7] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *CVPR*, 2005.

[8] O. Chum and J. Matas. Optimal randomized RANSAC. *IEEE Trans. PAMI*, 30(8):1472–1482, 2008.

[9] O. Chum, J. Matas, and J. Kittler. Locally optimized RANSAC. In *DAGM-Symposium*, pages 236–243, 2003.

[10] O. Chum, T. Werner, and J. Matas. Epipolar geometry estimation via RANSAC benefits from the oriented epipolar constraint. In *ICPR*, 2004.

[11] O. Chum, T. Werner, and J. Matas. Two-view geometry estimation unaffected by a dominant plane. In *CVPR*, 2005.

[12] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24(6), 1981.

[13] J.-M. Frahm and M. Pollefeys. RANSAC for (quasi-) degenerate data (QDEGSAC). In *CVPR*, pages 453–460, 2006.

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

[15] K.-M. Lee, P. Meer, and R.-H. Park. Robust adaptive segmentation of range images. *IEEE PAMI*, 20:200–205, 1998.

[16] J. Meidow, C. Beder, and W. Förstner. Reasoning with uncertain points, straight lines, and straight line segments in 2d. *ISPRS JPRS*, 64(2):125–139, 2009.

[17] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[18] J. Miller and C. V. Stewart. MUSE: Robust surface fitting using unbiased scale estimates. In *CVPR*, 1996.

[19] K. Ni, H. Jin, and F. Dellaert. GroupSAC: Efficient consensus in the presence of groupings. In *ICCV*, 2009.

[20] D. Nistér. Preemptive RANSAC for live structure and motion estimation. In *ICCV*, 2003.

[21] R. Raguram, J.-M. Frahm, and M. Pollefeys. A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In *ECCV*, 2008.

[22] R. Raguram, J.-M. Frahm, and M. Pollefeys. Exploiting uncertainty in random sample consensus. In *ICCV*, 2009.

[23] P. J. Rousseeuw. Least median of squares regression. *Journal of the ASA*, 79(388):871–880, 1984.

[24] T. Sattler et al. SCRAMSAC: Improving RANSAC's efficiency with a spatial consistency filter. In *ICCV*, 2009.

[25] M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the ASA*, 69(347):pp. 730–737.

		Line		**Plane**		**H**		**F**	
		k	err	k	err	k	err	k	err
$\varepsilon = 0.6$ $\sigma = 3.0$	RANSAC$_F$	106.8	1.38	137.1	1.55	218.3	3.33	266.2	3.58
	LMedS$_F$	11	1.24	23	1.38	47	2.9	382	2.98
	RANSAC$_T$	23.1	1.23	31.5	1.41	66.4	2.97	271.1	3.1
	LMedS$_T$	11	1.24	23	1.38	47	2.9	382	2.98
	RECON	29.6	1.2	38.2	1.35	49.4	2.91	133.1	3.04
$\varepsilon = 0.4$ $\sigma = 2.0$	RANSAC$_F$	176.2	0.91	389.5	1.2	489.2	3.17	3544.4	2.98
	LMedS$_F$	11	0.87	23	1.41	47	9.31	382	12.21
	RANSAC$_T$	56.4	0.7	99.7	0.88	304.7	2.05	2431.5	2.28
	LMedS$_T$	47	0.72	191	0.8	766	2	49082	1.88
	RECON	46.9	0.73	72.5	0.89	166.9	2.04	1879.3	2.34
$\varepsilon = 0.3$ $\sigma = 2.0$	RANSAC$_F$	289.8	1.08	629.2	1.36	1101.8	3.18	23450.5	3.26
	LMedS$_F$	11	2.01	23	5.72	47	17.2	382	25.33
	RANSAC$_T$	79.7	0.7	212	0.82	891.2	2.55	30153.5	2.42
	LMedS$_T$	47	0.77	191	0.74	766	2.51	49082	2.13
	RECON	58.3	0.75	149.5	0.76	490	2.56	13751.4	2.44
$\varepsilon = 0.3$ $\sigma = 4.0$	RANSAC$_F$	344.6	1.62	702.3	2.49	1798.3	6.03	94303.3	5.39
	LMedS$_F$	11	2.85	23	6.12	47	24.4	382	46.45
	RANSAC$_T$	115.1	0.96	291.8	1.67	1373.4	4.61	35983.9	4.06
	LMedS$_T$	47	1.03	191	1.65	766	4.48	49082	3.61
	RECON	77.2	0.97	172.4	1.72	612.3	4.47	13894.4	3.98

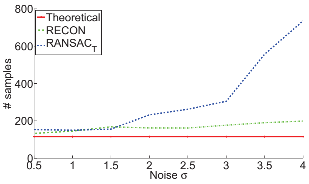


Fig. 4(a)

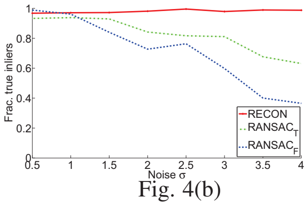


Fig. 4(b)

[Left: Table 1] Results on synthetic data. The table lists the number of samples (k) and mean error (err) for each dataset and algorithm. err is the mean error of ground truth inliers to the recovered model (perpendicular dist. for line and plane, symmetric transfer error for **H**, Sampson error for **F**). RANSAC$_T$/LMedS$_T$ refer to running the algorithms with ground truth parameters, while RANSAC$_F$/LMedS$_F$ refer to fixed parameters. Note that RECON provides the same accuracy as RANSAC$_T$/LMedS$_T$, without any *a priori* knowledge of the parameters. **[Right]** Plots for homography estimation with $\varepsilon = 0.4$. **Fig 4(a)** Number of samples vs. σ. Note that RECON is significantly close to the theoretical estimate. **Fig 4(b)** Fraction of true inliers recovered vs. σ. RECON returns close to 100% of the inliers.

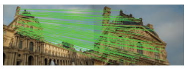
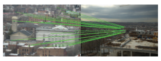
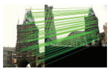
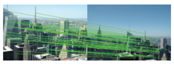
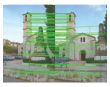
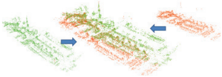
		RANSAC	RECON			RANSAC	RECON
Homography (**H**)				Essential matrix (**E**)			
 a : $N = 2450$	k I err	66.5 ± 21 1344 ± 96 2.63 ± 0.53	37.9 ± 7 1663 ± 38 2.62 ± 0.52	 **g** : $N = 962$	k I err	72.3 ± 25 556 ± 21 1.29 ± 0.61	58.9 ± 15 598 ± 14 1.25 ± 0.63
 **b** : $N = 495$	k I err	588.3 ± 154 153 ± 9 2.59 ± 0.49	224.6 ± 43 174 ± 3 2.59 ± 0.46	 **h** : $N = 1110$	k I err	108.2 ± 31 602 ± 33 0.69 ± 0.21	78.5 ± 19 632 ± 25 0.68 ± 0.23
 c : $N = 1967$	k I err	256.6 ± 121 739 ± 52 1.47 ± 0.45	73.3 ± 22 995 ± 23 1.47 ± 0.42	 **i** : $N = 1207$	k I err	1092.3 ± 312 417 ± 41 1.72 ± 0.77	407.4 ± 57 465 ± 20 1.72 ± 0.74
Fundamental matrix (**F**)				3D similarity			
 d : $N = 2454$	k I err	52.3 ± 26 1854 ± 33 1.30 ± 0.85	39.1 ± 16 1897 ± 28 1.31 ± 0.86	 **j** : $N = 4529$	k I err	78.1 ± 25 1640 ± 27 $0.09 \pm 2\text{x}10^{-3}$	59.9 ± 8 1887 ± 14 $0.08 \pm 1\text{x}10^{-3}$
 e : $N = 930$	k I err	11514.3 ± 2213 311 ± 8 1.91 ± 0.72	4065.2 ± 565 341 ± 4 1.88 ± 0.69	 **k** : $N = 1955$	k I err	43.3 ± 12 967 ± 36 $8.2\text{x}10^{-3} \pm 2\text{x}10^{-4}$	31.1 ± 8 982 ± 21 $6.7\text{x}10^{-3} \pm 5\text{x}10^{-4}$
 **f** : $N = 838$	k I err	148.8 ± 48 520 ± 30 1.78 ± 0.83	72.5 ± 23 568 ± 18 1.8 ± 0.86	 **l** : $N = 2549$	k I err	140.8 ± 28 825 ± 51 $0.02 \pm 3.1\text{x}10^{-3}$	99.0 ± 12 $844{\pm}20$ $0.02 \pm 1.0\text{x}10^{-3}$

Table 2. Results on real data, for four estimation problems: homography (**a**-**c**), fundamental matrix (**d**-**f**), essential matrix (**g**-**i**) and 3D similarity (**j**-**l**). The table lists the number of hypotheses generated (k), the number of inliers returned (I), the total number of points (N), and the mean error (err). err is the mean over inliers with respect to the returned model (symmetric transfer error for **H** and 3D similarity, Sampson distance for **F** and **E**). Units are in pixels for **H**, **F** and **E**. Note that the error for 3D similarity is arbitrarily scaled. By using real world measurements, the error in **l** was found to correspond to ≈ 8 cm.

[26] C. V. Stewart. MINPRAN: A new robust estimator for computer vision. *IEEE Trans. PAMI*, 17(10):925–938, 1995.

[27] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *IJCV*, 24:271–300, 1997.

[28] H. Wang and D. Suter. Robust adaptive-scale parametric model estimation for computer vision. *IEEE Trans. PAMI*, 26(11):1459–1474, Nov. 2004.

[29] W. Zhang and J. Kosecka. A new inlier identification procedure for robust estimation problems. In *RSS*, 2006.