Model Selection and Surface Merging in Reconstruction Algorithms

Kishore Bubna[1]
bubnak@cs.rpi.edu

Charles V. Stewart
stewart@cs.rpi.edu

Department of Computer Science
Rensselaer Polytechnic Institute
110, 8th Street
Troy, NY 12180-3590

[1]The authors would like to thank James Miller for helpful discussions and comments. They would also like to acknowledge the financial support of the National Science Foundation under grant IRI-9408700.

Abstract

The problem of model selection — automatically choosing the correct function to describe a data set — is relevant to many areas of computer vision. Many model selection criteria have been used in the vision literature and many more have been proposed in statistics, but the relative strengths of these criteria have not been analyzed in vision. Using the problem of surface reconstruction as our context, we analyze existing criteria using simulations and real data, introduce new criteria from statistics, develop novel criteria capable of handling unknown error distributions and outliers, and extend model selection criteria to apply to the surface merging problem. The new surface merging rules improve upon previous results, and work well even at small step heights ($h = 3\sigma$) and crease discontinuities. Our results show that when the error distribution is known (at least approximately), Bayesian criteria for model selection and surface merging introduced here works best, although for time-sensitive applications a variant of the Akaike criterion may be a better choice. For unknown distributions, the Bayesian criteria combined with a bootstrapped estimate of the error distribution gives the best performance. Unfortunately, none of the criteria work reliably for small datasets, implying that model selection and surface merging should be avoided unless there is sufficient data.

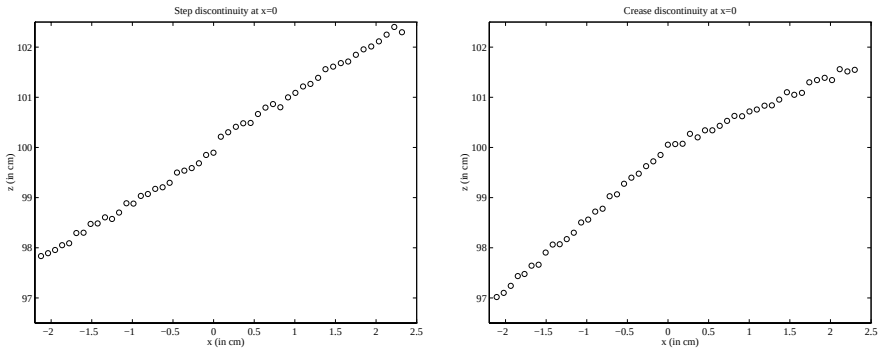
1 Introduction

Determining the correct function (or model) to describe a data set is an important issue in computer vision, arising in a variety of domains such as segmentation, surface reconstruction, 3D modeling, object recognition, reverse engineering, inspection, and tracking. In each of these domains, the appropriate model must be chosen automatically if it is not dictated by prior knowledge. This "model selection" problem has received much less attention than the associated problem of estimating model parameters [8, 9, 32, 33, 39], yet without a proper model, the estimated parameters have little meaning.

The model selection criteria that have been used in vision have many origins. Many of these criteria are heuristic [6, 8, 36, 43] and some rely on user defined thresholds. Others, especially recent ones, are applications of statistical and information theoretic criteria [7, 9, 16, 27, 29, 30, 41, 46, 47]. Unfortunately, the advantages and limitations of these criteria in computer vision algorithms have not been carefully analyzed. Most do not work well for small region sizes, most make errors near small magnitude discontinuities, and some are biased towards higher or lower order models. Further, most model selection criteria have been derived to choose a single model — e.g. a planar, quadratic, or higher order model — for a given set of data, but some vision problems require criteria that can decide between describing a data set with a single model or partitioning the data set and describing each with a separate model. Finally, model selection criteria in vision must tolerate outliers [9, 16], unknown noise distributions, and other kinds of unmodeled errors in the data [15].

One computer vision problem where model selection techniques are crucial is surface reconstruction. Many reconstruction algorithms use a local-to-global approach in which parameter estimation techniques and local decision criteria are combined in a greedy surface recovery strategy. This approach generally consists of two steps. In the first step, initial surface patches are fit to the data either by dividing the image into a grid [13, 44], by clustering methods [22, 36] or by region growing [6, 9, 29, 43]. Model selection is an important part of this initial estimation step. For example, when expanding "seed regions", at each iteration, it must be decided whether to continue growing using the same model or to switch to a different model. This difficult problem has been addressed in vision with a variety of techniques, including heuristic criteria based on user-defined thresholds [6, 29, 43], Chi-square tests [46], runs test [6, 8], Bayes rule [7, 14, 41, 47], Kullback-Liebler (K-L) distance [9], and minimum description lengths (MDL) [16, 27, 29, 30].

Depending on the algorithm, this first, "surface growing" step of the local-to-global reconstruction process either gives redundant surfaces [9, 29], or creates artificial surface boundaries [6, 13, 36, 43, 44]. Therefore, in the second step, surface patches need to be *pruned* or surface



(a) (b)

Figure 1: Existing techniques tend to identify each of the above set of points as a single linear or quadratic surface. However, (a) shows a step discontinuity at $h = 4\sigma$ and (b) shows a crease discontinuity in which the two line segments make an angle of 160 degrees. The points were generated using our sensor model (see Section 6) at $\sigma = 0.05$ cm.

patches need to be *merged* to form larger surfaces. Pruning techniques combine model selection criteria with greedy search techniques, and their design must balance between leaving multiple surface descriptions for some data sets and making others devoid of any surface description. Merging decisions, on the other hand, must balance between bridging true surface boundaries and leaving artificial surface boundaries, a difficult problem illustrated by the examples shown in Figure 1. In an attempt to avoid the model selection problem, many merging techniques only join surfaces fit to the same model [6, 36, 44, 13, 26], potentially limiting the effectiveness of the merging process.

Surface reconstruction, therefore, provides a good context for studying the model selection problem in computer vision. The effectiveness of different selection criteria in surface reconstruction has not been studied carefully, the problem of merging surfaces over different models continues to be difficult, and many existing criteria do not account for outliers and unknown noise distributions, common in range data. To alleviate these problems, we study the characteristics of different model selection criteria used in the vision and statistics literature, modify them for use in the presence of outliers, develop new criteria based on bootstrapped data distributions which do not require a prior model of the noise distribution, and finally, extend model selection criteria to develop new techniques for surface merging. All new and existing criteria studied in this paper are free from user-defined, data dependent, thresholds, although some use statistical thresholds (confidence intervals). We compare the relative performance of these criteria using simulated data (containing small-scale Gaussian errors), and real data (containing small-scale random errors and outliers). Our results show that many of these new criteria can be used to give improved performance over existing techniques (for example, the discontinuities

in Figure 1 can be detected by using new techniques presented in this paper). The experiments determine the performance of these criteria under different conditions, and identify situations in which they perform poorly. These results, therefore, can be used to decide among different model selection and merging criteria for different types of data and applications.

2 Definitions

This section defines the terms and concepts used in the rest of the paper.

Range image: A range image is characterized by a point $\boldsymbol{p}_i = [\boldsymbol{x}_i \; z_i]^T$ at any pixel i in the image. For our simulations, $\boldsymbol{x}_i$ will simply be a scalar x_i, and for real range images $\boldsymbol{x}_i = [x_i \; y_i]^T$. We call the former 2D range images and the latter 3D range images. For this paper, we assume errors in range data are all in the depth (z) direction[1].

Candidate models and orthogonal polynomials: There are two practical problems associated with model selection. First, when fitting higher order models to data belonging to a low order surface, the design matrix may be ill-conditioned. Second, fitting different models to a data set is computationally expensive. To alleviate these problems, we fit using orthogonal polynomials [3, 4]. Least-squares estimation using orthogonal polynomials gives well-conditioned matrices, and is efficient because the fit to high order models builds on fits to lower order models. When using robust estimation techniques, the second advantage is lost because outliers are determined differently for different models, so that parameters must be estimated separately for each model order.

Experiments in this paper are based on data sets from linear and quadratic models. To test performance of different criteria we use the set $M = \{m_0, m_1, m_2, m_3\}$ of candidate models, where m_0 stands for the zeroth order model and m_3 for the cubic model. m_0 and m_3 are included in M to detect bias toward low or high order models in different criteria. The models in M use discrete orthogonal polynomials as basis functions [3, 5], and are given by

$$z(\boldsymbol{x}) = \sum_{j=0}^{d_m - 1} \theta_j \; \phi_j(\boldsymbol{x}), \quad m = 0 \ldots 3, \tag{1}$$

where d_m is the number of parameters in the model. The parameter vector then is given by $\boldsymbol{\theta}_m = [\theta_0 \; \theta_1 \; \ldots \theta_{d_m - 1}]^T$. The set of orthogonal basis polynomials, $\phi_j(\boldsymbol{x})$, is constructed using

[1]Errors in sensor data are, generally, along all coordinate directions. But, experiments with our sensors show that for relatively small fields of view (a viewing cone of 25 degrees or less), the errors can be approximated to be along the depth (z) direction. Almost all other algorithms assume the same [2], and have not reported any problems.

the n data points, and satisfies the relation[2]

$$\sum_{i=1}^{n} \phi_p(\boldsymbol{x}_i)\, \phi_q(\boldsymbol{x}_j) = 0, \quad \text{for } p \neq q. \tag{2}$$

Parameter estimation for model selection: Model selection techniques fit each of the candidate models to the data, basing the choice of models on various measures of fit accuracy and model complexity. Hence, parameter estimation forms an important part of model selection.

Let $D = \{\boldsymbol{p}_1, \ \ldots\ , \ \boldsymbol{p}_n\}$ be a set of n data points. Consider models of the form

$$\boldsymbol{Z} = \boldsymbol{X}_m\, \boldsymbol{\theta}_m \ + \ \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{Z}$ contains the $(n \times 1)$ depth values, $\boldsymbol{X}_m$ contains $(n \times d_m)$ orthonormal polynomials where any element $\boldsymbol{X}_m(i,j) = \phi_j(\boldsymbol{x}_i)$, and $\boldsymbol{e} = [e_1\ e_2\ \ \ldots e_n]^T$ is a vector of unobserved, but independent random variables. Since many model selection criteria are based on the loglikelihood of the estimated parameters, maximum likelihood estimators must be used. We use ordinary least-squares for data with Gaussian errors, and M-estimators, solved using iteratively reweighted least squares (IRLS) [23], for data with random errors and outliers. Only M-estimators that are maximum likelihood estimators [20, page 361-362] can be used in model selection. Following Boyer, Mirza, and Ganguly [9], we use the M-estimator based on a t-distribution.

The loglikelihood when $\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ and when σ is known *a priori* is

$$\log L(\boldsymbol{\theta}_m) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \left(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_m\right)^T \left(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_m\right). \tag{4}$$

Maximization yields the estimate $\hat{\boldsymbol{\theta}}_m$. When σ in (4) is not known, the log likelihood is written as $\log L(\boldsymbol{\theta}_m, \sigma_m)$ and maximization yields both $\hat{\boldsymbol{\theta}}_m$ and $\hat{\sigma}_m$. In the presence of outliers, the errors are reasonably represented by a t-distribution [9], having degree of freedom f and scaled by σ, giving

$$\log L(\boldsymbol{\theta}_m) = -\frac{1+f}{2} \sum_{i=1}^{n} \ w_{mi} \ \log\left(1 + \frac{e_{mi}^2}{f\sigma^2}\right), \quad w_{mi} = \frac{1+f}{f + (e_{mi}/\sigma)^2}. \tag{5}$$

Here, e_{mi} is the error for the ith observation. $\hat{\boldsymbol{\theta}}_m$ is estimated from (5) using IRLS initialized by the least median of squares (LMS) [32] estimate of $\boldsymbol{\theta}_m$. When σ in (5) is unknown, the log likelihood is written as $\log L(\boldsymbol{\theta}_m, \sigma_m)$ and $\hat{\sigma}_m$ is estimated using IRLS (see [9] for details).

[2] Besl [4, 5] derived orthogonal polynomials by assuming that $\boldsymbol{x}_1, \ \ldots\ , \ \boldsymbol{x}_n$ are equally spaced. For perspective projection measurements this assumption is violated, and we derive orthogonal polynomials using the method given in [3].

3 Model selection

This section describes model selection criteria already used in reconstruction algorithms and introduces several promising criteria from the statistics literature. Some of these are extended to work with bootstrapped error distributions in Section 4.

3.1 Information-theoretic criteria

Model selection using K-L distance: Some of the earliest criteria select the model minimizing the Kullback-Leibler (K-L) distance:

$$d(\hat{\boldsymbol{\theta}}_m, \boldsymbol{\theta}_*) = E_*[-2 \log L(\boldsymbol{\theta}_m)]|_{\boldsymbol{\theta}_m = \hat{\boldsymbol{\theta}}_m}, \tag{6}$$

where $\boldsymbol{\theta}_*$ represents the parameters for the "true" or *generating model* and E_* denotes the expectation under the generating model. The Akaike Information Criterion (AIC) [1] approximates (6) by

$$d(\hat{\boldsymbol{\theta}}_m, \boldsymbol{\theta}_*) \approx -2 \log L(\hat{\boldsymbol{\theta}}_m) + 2d_m. \tag{7}$$

While AIC has not been used in surface reconstruction, [9] uses a popular variant of AIC, CAIC [10]

$$d(\hat{\boldsymbol{\theta}}_m, \boldsymbol{\theta}_*) \approx -2 \log L(\hat{\boldsymbol{\theta}}_m) + d_m(\log n + 1). \tag{8}$$

We study both CAIC and AIC here. When σ is unknown, $L(\hat{\boldsymbol{\theta}}, \hat{\sigma})$ replaces $L(\hat{\boldsymbol{\theta}})$ in (7) and (8).

Model selection using Bayes rule: Criteria based on Bayes rule choose the model that maximizes the probability of the data, D, given the model m and prior information I. This probability is denoted by $P(D|m, I)$. Using Bayes rule (and assuming $\boldsymbol{\theta}_m$ and σ are independent [31, page 109]),

$$P(D|m, I) = \int \int \; P(D|\boldsymbol{\theta}_m, \sigma, m, I) \; P(\boldsymbol{\theta}_m|m, I) P(\sigma|I) \; d\boldsymbol{\theta}_m d\sigma \tag{9}$$

$P(D|\boldsymbol{\theta}_m, \sigma, m, I)$ in (9) is just the likelihood $L(\boldsymbol{\theta}_m)$. $P(\boldsymbol{\theta}_m|m, I)$ is the prior probability of $\boldsymbol{\theta}_m$. Since reconstruction applications generally lack prior information on parameters, we use a uniform prior on $\boldsymbol{\theta}_m$ (see [21, appendix A]). When σ is known, its prior, $P(\sigma|I)$, is a delta function at the known σ, and (9) reduces to an integral with respect to $\boldsymbol{\theta}_m$ only. Solving this reduced integral using a second order Taylor's expansion of $\log L(\boldsymbol{\theta}_m)$ at $\hat{\boldsymbol{\theta}}_m$ [25, chapter 24] yields

$$P(D|m, I) \approx (2\pi)^{d_m/2} L(\hat{\boldsymbol{\theta}}_m)[|-\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m)|]^{-1/2}, \tag{10}$$

where $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m)$ is the Hessian of $\log L(\boldsymbol{\theta}_m)$ at $\hat{\boldsymbol{\theta}}_m$. When σ is not known, we need to assign $P(\sigma|I)$. Again, we use non-informative priors on $P(\sigma|I)$. For the Gaussian case, using the non-informative

prior $1/\sigma$ for σ (see [25, chapter 6, page 29]), and assigning other probabilities as before, (9) reduces to

$$P(D|m,I) = \frac{\Gamma((n-d_m)/2)}{2^{(d_m/2)+1}\ \pi^{n/2}\ |\boldsymbol{X}_m^T\boldsymbol{X}_m|^{1/2}\ RSS_m^{(n-d_m)/2}}, \tag{11}$$

where RSS_m is the residual sum of squares for model m. For t-distributed errors, assuming a uniform prior on σ [21], (9) reduces to

$$P(D|m,I) = (2\pi)^{d_m/2}\ L(\hat{\boldsymbol{\theta}}_m, \hat{\sigma})\ [|-\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m, \hat{\sigma})|]^{-1/2}, \tag{12}$$

where $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m, \hat{\sigma})$ is the Hessian of $\log L(\boldsymbol{\theta}_m, \sigma)$ at the maximum likelihood estimates $\hat{\boldsymbol{\theta}}_m$ and $\hat{\sigma}_m$. These criteria, (10), (11) and (12), will be referred to as BAYES.

To avoid the expense of computing $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m, \hat{\sigma})$, several asymptotic approximations of (10) have been introduced. A common one is BIC [38]:

$$P(D|m,I) \approx L(\hat{\boldsymbol{\theta}}_m)\ n^{-d_m/2}, \tag{13}$$

(Once again $L(\hat{\boldsymbol{\theta}}_m, \hat{\sigma}_m)$ replaces $L(\hat{\boldsymbol{\theta}}_m)$ in (13) when σ is unknown.)

None of these Bayesian techniques have been used in surface reconstruction.

Model selection using MDL principle: The number of bits required to express the observed data using model m is $\mathrm{len}_m = \mathrm{len}(\hat{\boldsymbol{e}}_m) + \mathrm{len}(\hat{\boldsymbol{\theta}}_m)$, where len denotes the length of the bit string required to encode any quantity. Model selection criteria based on the MDL principle choose a model that minimizes len_m. The quantities $\mathrm{len}(\hat{\boldsymbol{e}}_m)$ and $\mathrm{len}(\hat{\boldsymbol{\theta}}_m)$ are calculated using different assumptions giving rise to different model selection criteria. The most common of these criteria is due to Rissanen [34], and is equivalent to BIC, equation (13). In [35], Rissanen derives an improved criterion which chooses the model minimizing

$$\mathrm{len}_m = -\log_2 L(\hat{\boldsymbol{\theta}}_m) + \frac{d_m}{2}\log_2{}^*\left(\hat{\boldsymbol{\theta}}_m^T\ (-H(\hat{\boldsymbol{\theta}}_m))\ \hat{\boldsymbol{\theta}}_m\right) + \log_2{}^* V_{d_m}, \tag{14}$$

where $\log_2{}^*(t) = \log_2 t + \log_2\log_2 t + \ldots$, including only its positive terms, and V_{d_m} is the volume of the d_m-dimensional unit hypersphere [17, page 24]. When σ is not known, (14) becomes

$$\mathrm{len}_m = -\log_2 L(\hat{\boldsymbol{\theta}}_m, \hat{\sigma}) + \frac{d_m}{2}\log_2{}^*\left([\hat{\boldsymbol{\theta}}_m^T\ \hat{\sigma}](-H(\hat{\boldsymbol{\theta}}_m, \hat{\sigma}))\ [\hat{\boldsymbol{\theta}}_m^T\ \hat{\sigma}]^T\right) + \log_2{}^* V_{d_m}. \tag{15}$$

Here $H(\hat{\boldsymbol{\theta}}_m, \hat{\sigma})$ is the Hessian of $\log_2 L(\boldsymbol{\theta}_m, \sigma)$ at the maximum likelihood estimates. Criteria (14) and (15), later referred as RISS, have not been used in the vision literature.

In surface reconstruction, Leonardis, Gupta and Bajcsy [29] uses a MDL based criterion in a quadratic optimization function. This criterion chooses a model that maximizes

$$K_1 n - K_2\ RSS_m - K_3 d_m, \tag{16}$$

where K_1, K_2, and K_3 are the average number of bits required to encode a data point, fit accuracy, and a fit parameter, respectively. The second term in (16), $-K_2\ RSS_m$, is equivalent to $\log_2 L(\hat{\boldsymbol{\theta}}_i)$ (see [28, page 65]), yielding

$$K_1 n + \log_2 L(\hat{\boldsymbol{\theta}}_m) - K_3 d_m \tag{17}$$

For model selection using a fixed data set, n is also fixed, allowing (17) to be simplified to

$$\log_2 L(\hat{\boldsymbol{\theta}}_m) - K_3 d_m \tag{18}$$

Although theoretically K_3 is the number of bits required to encode d_m, in practice the algorithm in [29] chooses the value of K_3 empirically. To compare the performance of (18) with other criteria, we need to choose a value for K_3. In this regard, comparing (18) with AIC (7), we see that when $K_3 = 1/\log 2$, maximizing (18) is equivalent to minimizing (7). Thus, the model selection criterion used in [29] is likely to show similar properties as that obtained using AIC. As such, we use AIC to study the properties of (16).

3.2 Model selection using Chi-square, F, and runs test

A number of model selection criteria that have been used in reconstruction algorithms are based on hypothesis tests. This section summarizes four such criteria. Each starts with the zeroth order model as the null hypotheses and moves to the next higher order model when a null hypotheses is rejected. In these techniques, since all null hypotheses may be rejected, it is possible that no model is selected.

This section also introduces a simple, new F-test model selection criteria. This technique may be used when σ is unknown and the Chi-square based techniques cannot be used.

RUNS: The intuition behind using a runs test is that low order incorrect models will produce a large "run" (consecutive sequence) of all positive or all negative residuals. For 2D range images, the total number of runs, r_m, for any fit $\hat{\boldsymbol{\theta}}_m$, is asymptotically[3] normally distributed and is given by [11, pages 164-170]

$$r_m \sim N\left(\frac{2p_m q_m}{p_m + q_m} + 1,\ \ \frac{2p_m q_m (2p_m q_m - p_m - q_m)}{(p_m + q_m)^2 (p_m + q_m - 1)}\right).$$

Here, p_m is the number of positive residuals, and n_m is the number of negative residuals in the fit. The test rejects model m if r_m is not within a 95% level of confidence. Since the RUNS test does not generalize to 3D range images, Besl [4, pages 150-152] introduces a heuristic approximation. He creates binary images of positive and negative residuals, erodes the images using a 3×3 kernel, finds the largest connected component in each image, and rejects the null hypotheses

[3]For small samples, techniques from [42] and [19] may be used.

if the larger of these components is greater than 2% of n. We follow this heuristic for 3D range images. The runs test is advantageous when σ and the noise distribution of the data are unknown.

CHI: This test is based on a one-way Chi-square test and rejects model m at a 95% confidence level. It has been used by Whaite and Ferrie [46]. The intuition is that low order incorrect models will produce a significant over-estimate to the error in the data.

BESL: This test combines RUNS and CHI, and rejects model m if both of them fail. This is the model selection criteria used by Besl and Jain [6].

RANSAC: This test [8] rejects model m for any one of three reasons: (a) CHI fails (this replaces the error-tolerance test using the RANSAC metric), (b) Reject m at 95% confidence level when $|p_m - n_m| > 2\sqrt{n}$, and (c) Reject m at 95% confidence level when the longest run exceeds $3.32 + \log_2 n$. For 3D range images, we replace the longest run with size of the largest connected component created by the process described in RUNS.

FTEST: In this test, any model m_i is rejected in favor of m_{i+1} if [45, page96]

$$\frac{(RSS_{m_i} - RSS_{m_{i+1}})/((n - d_{m_i}) - (n - d_{m_{i+1}}))}{RSS_{m_i}/(n - d_{m_{i+1}})} > F_{(d_{m_{i+1}} - d_{m_i}, n - d_{m_{i+1}});0.95}. \tag{19}$$

Starting with the zeroth-order model, this test continues switching to a higher order model until (19) is not satisfied or until all models in M have been tested.

4 New criteria for model selection using bootstrap principle

The model selection criteria presented thus far, with the exception of RUNS in Section 3.2, implicitly assume the error distribution is known *a priori*. Some, such as the techniques based on Chi-square tests and F test in Section 3.2 are more restrictive, specifically assuming a Gaussian distribution. In computer vision problems, however, error distributions are often unknown and difficult to model accurately. Because of this, it is crucial to develop model selection criteria that depend on only weak assumptions about error distributions. We address this problem in this section by deriving bootstrap [18] versions of the criteria in Section 3.1. The only assumptions are that the errors are zero-mean and independent. The resulting criteria are empirical in nature, making them somewhat expensive to compute, but they do not require user-defined thresholds and can be used when sensor error models are unavailable or unreliable.

The bootstrap is a method for estimating an unknown distribution from available data. Not yet popular in computer vision [12], this technique was introduced in statistics by Efron [18]. In linear regression, the bootstrap technique can be used to obtain an empirical distribution of errors in the data, and this distribution can then be used to generate different statistics on $\boldsymbol{Z}$ and $\hat{\boldsymbol{\theta}}_m$. The idea of bootstrap is simple. Consider the regression model given by (3), and let P be

the unknown distribution of the elements in $\boldsymbol{e} = [e_1 \dots e_n]^T$. Let $\hat{\boldsymbol{\theta}}_m$ be the least-square estimate of $\boldsymbol{\theta}_m$, let $\hat{\boldsymbol{e}}_m$ be the corresponding residuals, and let $\hat{Z}_m$ be the corresponding estimates of the uncorrupted Z. The residuals, $\hat{\boldsymbol{e}}_m = [\hat{e}_{m1} \dots \hat{e}_{mn}]^T$, can be used to generate an empirical distribution function, $\hat{P}_m$. $\hat{P}_m$ is defined to be the discrete distribution that puts probability $1/n$ on each value $\hat{e}_{m1}, \dots, \hat{e}_{mn}$. The plug-in bootstrap principle [18, chapter 4] substitutes P with $\hat{P}_m$ and this is used to generate R bootstrap error vectors, $\boldsymbol{e}^*_{1m}, \dots, \boldsymbol{e}^*_{Rm}$. Note that sampling from $\hat{P}_m$ is the same as sampling from the set $\{e_1, \dots, e_n\}$ with replacement. Each $\boldsymbol{e}^*_{im}$ is added to $\hat{Z}_m$ to generate a bootstrap set of $\boldsymbol{Z}$ values, $\boldsymbol{Z}^*_{im}$. These "bootstrap data" set can now be used to generate bootstrap least-square estimates $\hat{\boldsymbol{\theta}}^*_{1m}, \dots, \hat{\boldsymbol{\theta}}^*_{Rm}$.

Here, we derive bootstrap based versions of the model selection criteria described in Section 3.1. Each of these criteria is in the form of a penalized likelihood, balancing between accuracy and complexity of the model given the data. These two quantities need to be approximated using bootstrap data. First consider the accuracy term, which is given by the likelihood. When errors are unknown, ordinary least-squares is used for parameter estimation. More sophisticated estimators are unnecessary because our weak assumptions on sensor noise are sufficient to yield unbiased, minimum variance estimates [31, page 172]. The accuracy of the model given the data can then be measured using the normalized residual sum of squares, making prior knowledge of error distribution unnecessary. Therefore, we replace the model accuracy term $\log L(\hat{\boldsymbol{\theta}}_m)$ with $-RSS_m/\sigma^2$. In most vision applications, not only the error distribution, but σ is also unknown. To solve this problem, σ^* — the bootstrap estimate of σ — is calculated by finding the average standard deviation of $\boldsymbol{Z}^*_{1m}, \dots, \boldsymbol{Z}^*_{Rm}$. Unfortunately, this gives four different σ^*s corresponding to the four models in M. However, our experiments show that the σ^* values estimated using the correct model and those using any model of higher order than the correct model are close to each other and to the true σ[4]. This indicates that σ^*_3, which is the estimate from the cubic model, can be used for σ^* in $-RSS_m/\sigma^2$.

$\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m)$ is also needed for BAYES (10) and for RISS (14). To obtain a distribution-free measure of $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m)$, observe that $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m) \approx -[\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)]^{-1}$, the covariance matrix of $\hat{\boldsymbol{\theta}}_m$ [25, chapter 24]. This makes the problem easy because the bootstrap estimate of $\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)$, $\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)$ can be calculated from $\hat{\boldsymbol{\theta}}^*_{1m}, \dots, \hat{\boldsymbol{\theta}}^*_{Rm}$. Using this estimated covariance we obtain a bootstrapped, Bayesian model selection criterion (BMSC-BAYES) by taking the natural logarithm of (10), replacing $\log L(\hat{\boldsymbol{\theta}}_m)$ with $-RSS_m/{\sigma^*_{m_3}}^2$ and $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m)$ with $-[\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)]^{-1}$:

$$\text{BMSC-BAYES}_m = \frac{d_m}{2}\log 2\pi \; - \; \frac{RSS_m}{{\sigma^*_m}^2} \; + \; \frac{1}{2}\log|\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|. \tag{20}$$

[4]Note the same is not true for least-squares estimate of σ, but only the bootstrap estimates of σ, generated as explained above.

noise distribution	σ known	σ unknown
known (any)	AIC, BIC, CAIC, BAYES, RISS, RUNS	AIC, BIC, CAIC, BAYES, RISS, RUNS
known (Gaussian)	CHISQ, RANSAC, BESL, FTEST	FTEST
unknown	RUNS, BMSC-BAYES, BMSC-RISS	RUNS, BMSC-BAYES, BMSC-RISS

Table 1: Summary of model selection criteria.

Similarly, RISS (14) can be approximated as

$$\text{BMSC-RISS}_m = \frac{RSS_m}{\sigma_m^{*\,2}} \; + \; \frac{d_m}{2} \log_2 * \left(\hat{\boldsymbol{\theta}}_m^T \, |\boldsymbol{V}^*(\hat{\boldsymbol{\theta}}_m)|^{-1} \, \hat{\boldsymbol{\theta}}_m^T \right) \; + \; log_2 * (V_{d_m}). \tag{21}$$

These are the two bootstrap based criteria we study here. (Observe that while BMSC-BAYES needs to be maximized, BMSC-RISS needs to be minimized.) Since parameter estimation is based on least-squares, these criteria are true only for data with small-scale random errors. Similar bootstrap model selection criteria may be formulated for data with outliers, but this is relatively difficult and is part of our ongoing work.

See Table 1 for a summary of criteria presented in Section 3 and here.

5 New rules for surface merging

This section extends the model selection framework to develop new rules for merging surface patches to a single surface description. We assume for the discussion that small surface patches have already been estimated using different approaches discussed in Section 1, and these small surface patches do not undersegment the scene, i.e, they do not bridge discontinuities.

To define the problem precisely, suppose two surfaces A and B are fit to noisy data sets D_A and D_B where $D_A \bigcap D_B = \emptyset$. The issue is to determine whether D_A and D_B are measurements from the same or different underlying surfaces. When the surfaces are different, A and B should remain as fits to D_A and D_B. Conversely, when D_A and D_B are measurements from the same surface, they should be merged into a single surface, C, which can use any model $m \in M$. Let $C_0, \ldots, C_3$ be fits to the data set $D_C = D_A \bigcup D_B$, corresponding to models $m_0, \ \ldots \ , \ m_3$. Surface merging involves a choice between selecting $\{A, B\}$ for D_A and D_B or any one of C_0, C_1, C_2, and C_3 for D_C.

5.1 Formulation of rules based on information-theoretic criteria

As seen in Section 3.1, the different information theoretic criteria compare K-L distances, Bayesian probabilities, and minimum description lengths to select the best model from $m_0, \ldots \ m_3$. For

surface merging, we extend this notion, and compare the same quantities to select models m_A and m_B together or any one of models $m_0, \ldots, m_3$ for the data set $D_C = D_A U D_B$. Therefore, in addition to calculating the information-theoretic quantities for models $m_0, \ldots, m_3$, similar quantities need to be formulated for m_A and m_B combined. In this regard, note that since D_A and D_B are disjoint, their joint likelihood to fits from m_A and m_B are $L(\boldsymbol{\theta}_{m_A})L(\boldsymbol{\theta}_{m_B})$. The K-L distance of this likelihood under the generating model is then $E_*[-2\log L(\boldsymbol{\theta}_{m_A})L(\boldsymbol{\theta}_{m_B})]$ (see (6)), and evaluating at maximum likelihood estimates this reduces to $d(\hat{\boldsymbol{\theta}}_A, \boldsymbol{\theta}_*) + d(\hat{\boldsymbol{\theta}}_B, \boldsymbol{\theta}_*)$. Similarly, in the Bayesian case, $P(D_A D_B|m_A, m_B, I) = P(D_A|m_A, I)\ P(D_B|m_B, I)$, and in the MDL case $len_{m_A,m_B} = len_{m_A} + len_{m_B}$. Based on this, extended model selection may be represented as

$$\textbf{K-L distance}: \quad \min\{(d(\hat{\boldsymbol{\theta}}_A, \boldsymbol{\theta}_*) + d(\hat{\boldsymbol{\theta}}_B, \boldsymbol{\theta}_*)),\ d(\hat{\boldsymbol{\theta}}_{m_0}, \boldsymbol{\theta}_*),\ \ldots,\ d(\hat{\boldsymbol{\theta}}_{m_3}, \boldsymbol{\theta}_*)\}. \tag{22}$$

$$\textbf{Bayes rule}: \quad \max\{(P(D_A|m_A, I)\ P(D_B|m_B, I)),\ P(D|m_{m_0}, I),\ \ldots,\ P(D|m_{m_3}, I)\} \tag{23}$$

$$\textbf{MDL}: \quad \min\{(len_{m_A} + len_{m_B}),\ len_{m_1},\ \ldots,\ len_{m_3}\}, \tag{24}$$

Using (22) with (7) and (8) gives merging rules based on AIC and CAIC. Replacing (10), (11), and (12) in turn in (23) gives merging rules based on BAYES for σ known and unknown[5]. Similarly, using (14) or (15) with (24) gives merging rules based on RISS[6]. For formulating merging rules based on BMSC-BAYES and BMSC-RISS, note that these criteria are based on the logarithm of the Bayesian probability, $P(D|m, I)$, and RISS, respectively. As such, the respective merging rules correspond to (23) and (24), and are given by

$$\max\{(\text{BMSC-BAYES}_A + \text{BMSC-BAYES}_B),\ \text{BMSC-BAYES}_{m_0},\ \ldots,\ \text{BMSC-BAYES}_{m_3}\} \tag{25}$$

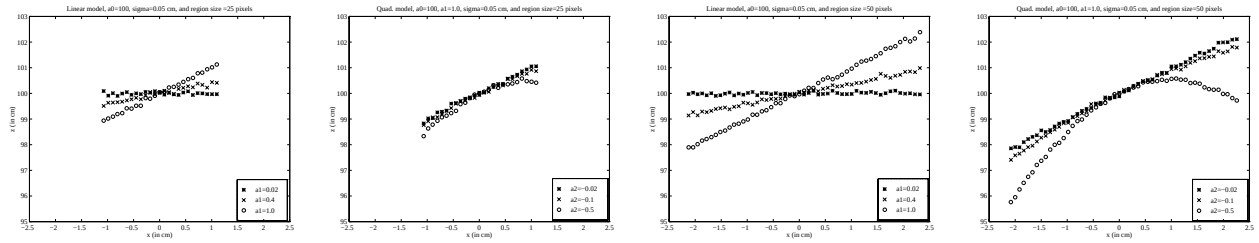
$$\min\{(\text{BMSC-RISS}_A + \text{BMSC-RISS}_B),\ \text{BMSC-RISS}_{m_0},\ \ldots,\ \text{BMSC-RISS}_{m_3}\}. \tag{26}$$

5.2 Formulation of rules based on statistical criteria

This section formulates simple merging rules using the model selection criteria discussed in Section 3.2. As mentioned before, tests RUNS to RANSAC in Section 3.2 may reject all four candidate models in M, and therefore, it is possible that no model is selected. Based on this observation, the rule merges A and B to C if a model from M is selected for C, otherwise, it concludes

[5] In surface reconstruction, a Bayesian merging approach has been used by LaValle and Hutchinson [26]. However, they only merge surfaces corresponding to the same model. They also restrict the parameter space such that $||\boldsymbol{\theta}_m|| = 1$. As such, their work can be considered as a special case of ours. Another difference is that they use a precise non-informative prior on $\boldsymbol{\theta}_m$, and solve the integral (9) numerically.

[6] Section 3.1 shows how the model selection criteria used in the optimization function in [29] is equivalent to AIC when $K_3 = 1/\log 2$. This function can be used for merging surfaces, and turns out to be equivalent to (22) used with AIC.



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

Figure 2: Plots (a) and (c) show data from linear model at different a_1, and (b) and (d) show data from quadratic model at different a_2. While (a) and (b) are at region size of 25 pixels, (c) and (d) are at 50 pixels. Note how data points marked '*' appear to be from a lower order model, those marked 'x' barely seem from the correct model, but points marked 'o' definitely seem to be from the correct model. The problem is more obvious at a region size of 25 pixels.

that there is a discontinuity. This technique can be applied to the first four criteria in Section 3.2 giving four different merging rules. Note how these do not use any information from the fitted surfaces A and B, but work directly on the data set D_C.

FTEST in Section 3.2 always selects a model, making the above technique inappropriate for surface merging. Our merging rule based on the F-test, therefore, works in two steps. In the first step, it checks if the parameters of surface A are within the 95% confidence interval of the parameters of B (or vice-versa — only one must succeed) [40] using the F statistic in [45, page 97]. When A and B belong to different models, the technique only checks if the lower order model fits within the confidence interval of the higher order model. If this step decides that the surfaces be merged, then the second step uses FTEST of Section 3.2 to find the best model.

6 Simulation results

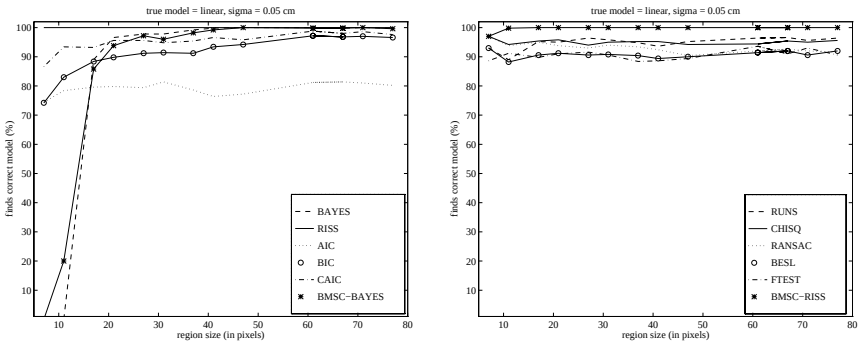
This section compares model selection criteria and merging rules on two-dimensional range images (see Section 2). The data contains Gaussian errors and are generated using focal length=1.77 cm and pixel size=0.0016 cm, the calibration parameters of our range sensor [37]. Simulated data allows us to test selection criteria on data from different region sizes and σ values, and to test merging rules under different step heights and crease angles. The rules and criteria assume Gaussian errors and known σ (except for those based on RUNS and bootstrap principle which make no assumption regarding the noise distribution or σ).

6.1 Model selection

The experiments are based on data sets from linear and quadratic models given by $z = a_0 + a_1 x$, and $z = a_0 + a_1 x + a_2 x^2$, respectively. Performance is compared at different region sizes and σ, and by varying a_1 for linear model, and a_2 for quadratic model (see Figure 2 for sample data at different region sizes, a_1 and a_2). Our sensor has a σ of about 0.02 cm at a depth of 100 cm, in our experiments we vary σ from 0.02 to 0.1 cm. The results are based on 500 simulations, and for bootstrap criteria, the number of bootstrap replications, R, is set to 200 [18].

Effect of region size and σ on performance: In this set of experiments, $a_0 = 100$ and $a_1 = 1$ (for both the models), and $a_2 = -0.1$ for the quadratic model. The experiment increases the region size symmetrically around the origin (see Figure 2) from 7 to 77 pixels and varies σ from 0.02 cm to 0.1 cm. Figure 3 shows percentage success of different selection criteria for data from the linear model at $\sigma = 0.05$ cm. Figure 3(a) shows results for information theoretic criteria and Figure 3(b) shows results for BMSC-RISS and the criteria based on confidence tests (Section 3.2). The results show that RISS performs the best, and although BAYES, BIC, and CAIC have problems at small region sizes, their performance improves as region size increases. BAYES performs the worst for small region sizes, but jumps to 97% success at a region size of about 20 pixels. The new bootstrap based criteria, BMSC-BAYES and BMSC-RISS also perform well and closely follow BAYES and RISS, respectively. This performance is promising, given that these criteria do not make any assumption regarding the noise distribution. For large region sizes, BAYES, RISS, and their bootstrap versions perform the best. This is closely followed by CAIC and BIC. The criteria based on significance tests have a success rate from 90 to 95%. This is expected because they are based on a 95% confidence interval. Surprisingly, however, AIC shows a success rate of only 80%, and tends to choose quadratic and cubic fits over a linear fit. Although not shown here, the results exhibit small improvements at $\sigma = 0.02$ and only minor performance hits at $\sigma = 0.1$.

Figure 4 shows corresponding performance at $\sigma = 0.05$ for data from the quadratic model (see Figure 2(b) and (d) for sample data). The results show that all criteria have problems at small region sizes, and show close to "steady state" performance (say, within 3% of maximum success rate) after a certain minimum region size. This minimum region size changes with σ. Table 2 shows the minimum region size for each criteria at different values of σ. The results show several differences from the linear case. First, with increasing σ, all criteria find it increasingly difficult to find the quadratic model at small region sizes (the relative performance of different criteria, however, remains similar to Figure 4). This is not surprising, given the difficulty in identifying a quadratic fit from the data in Figure 2(b). Second, RISS and BMSC-RISS, which perform the best for linear models even for small regions, now perform poorly at small region sizes. This suggests a possible bias in these criteria towards low order surfaces. But once again, at large re-



(a) (b)

Figure 3: Performance with increasing region size for data from linear model at $\sigma = 0.05$ cm.

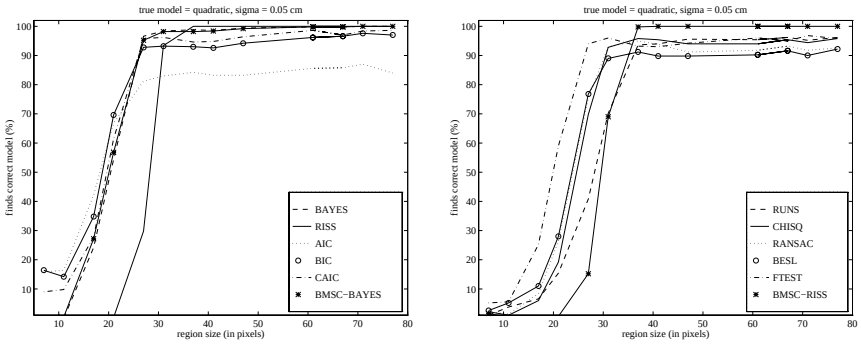


Figure 4: Model selection with changing region size for quadratic model at $\sigma = 0.05$ cm.

gion sizes BAYES, RISS, and their bootstrap versions perform the best. CAIC and BIC closely follow these criteria. AIC again shows a success rate of about 80%, and tends to choose cubic fits over a quadratic fit. Among the criteria based on significance tests, FTEST performs the best, RUNS matching its performance only at large region sizes. CHI, RANSAC, and BESL exhibit average performance. Comparing the information theoretic criteria and those based on confidence intervals, the former performs better at small region sizes. This is because information theoretic criteria are based on absolute comparisons and overcome the disadvantages of setting tolerance limits *a priori*.

Effect of changing a_1 and a_2: In this set of experiments, we vary a_1 for the linear model (with a_0 fixed at 100), and vary a_2 for the quadratic model (with a_0 fixed at 100 and a_1 fixed at 1) at $\sigma = 0.05$ cm and a region size of 25 pixels. Figure 5 shows the results for a linear model when varying a_1 from 0.02 to 0.4, and Figure 6 shows the results for a quadratic model when varying a_2 from -0.02 to -0.5. We see that all criteria find it difficult to choose the correct model when

criteria	min. region size (pixels)			criteria	min. region size (pixels)		
	$\sigma = 0.02$	$\sigma = 0.05$	$\sigma = 0.1$		$\sigma = 0.02$	$\sigma = 0.05$	$\sigma = 0.1$
BAYES	25	30	40	RUNS	30	40	50
RISS	25	35	45	CHI	25	35	45
AIC	25	25	35	RANSAC	20	30	45
BIC	25	35	40	BESL	25	35	45
CAIC	25	30	40	FTEST	20	30	35
BMSC-BAYES	25	30	40	BMSC-RISS	25	35	45

Table 2: For data from quadratic model, table shows the minimum region sizes required by model selection criteria at different values of σ to perform within 3% of their maximum success rate.

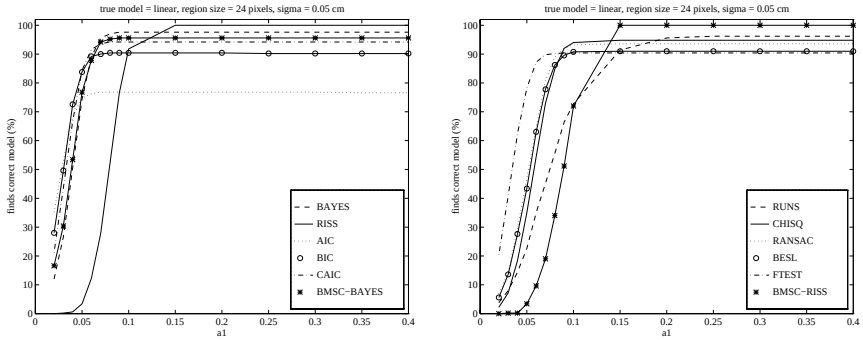


Figure 5: Model selection with changing a_1 for data generated from linear model.

a_1 and a_2 are small in magnitude (refer Figure 2(a) and (b)). Among these, RUNS, RISS, and BMSC-RISS are the most affected. However, RISS and BMSC-RISS perform the best when a_1 and a_2 are of relatively large magnitude. BAYES, BIC, and CAIC perform reasonably well, but BAYES again outperforms the other two at large magnitudes of a_1 and a_2. Among the confidence interval based criteria, FTEST performs better than the rest. Most criteria again reach the 90% to 100% success rate for larger magnitudes of a_1 and a_2. The bootstrap criteria continue to perform well, closely following their non-bootstrap versions, and AIC again shows poor performance.

To summarize then, most criteria perform well at moderate region sizes (greater than 25 pixels) under moderate noise levels ($\sigma \sim 0.05$ cm), and all criteria have problems at small region sizes, high values of σ, and at low magnitudes of a_1 and a_2. The results confirm intuition and match our own ability in detecting models from sample data in Figure 2. Note that the notion of a moderate region size of 25 pixels does not generalize to a 5x5 window in the 3D case. This point is further demonstrated by experiments in the next section. As far as specific criteria are concerned, AIC shows a bias towards higher order surfaces, while RISS and RISS-BMSC show a bias towards lower order surfaces. BAYES, CAIC, and BIC perform reasonably well, and

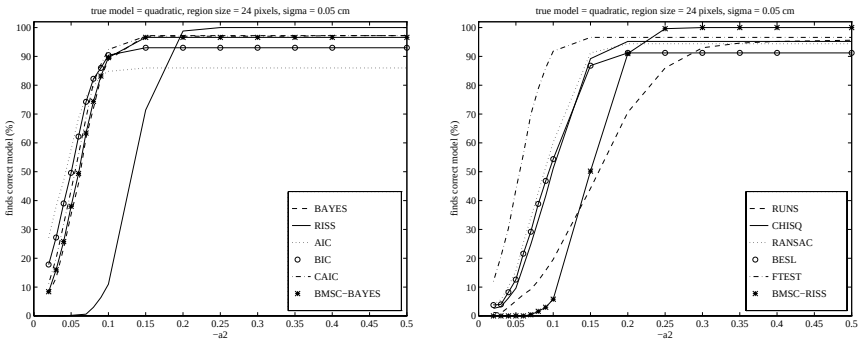


Figure 6: Model selection with changing a_2 for data generated from quadratic model.

Overall Performance	model selection criteria
Good	BAYES, BMSC-BAYES, CAIC, BIC, FTEST
Average	RISS, BMSC-RISS, RUNS, CHI, RANSAC, BESL
Poor	AIC

Table 3: Overall performance of model selection criteria with data generated using linear and quadratic models, and Gaussian errors.

BAYES outperforms the other two at larger region sizes. BMSC-BAYES introduced in this paper, performs as well as BAYES though it does not assume any noise distribution for the data. The confidence interval based criteria perform relatively worse at small region sizes, and reach a 90 to 95% success rate at larger region sizes. Thus, although there is no definite choice among the different criteria, it is clear that some criteria are preferable over others. Based on these results, Table 3 gives a qualitative summary of relative performance.

6.2 Surface Merging

This section compares the performance of different merging rules introduced in Section 5 on surface fits with step and crease discontinuities (see Figure 7), and artificial discontinuities (formed when $h = 0$ or $\alpha = 0$). The experiments are based on data generated from linear models.

For step discontinuities, data are generated from the following two surfaces:

$$A : z = (100 - \frac{h}{2}) + x, \quad B : z = (100 + \frac{h}{2}) + x.$$

Thus, A and B are separated by a step height of h cm. Figure 8 shows the percentage success of the merging rules in detecting a discontinuity at different values of h/σ at a region size of 25 pixels. The results show that merging rules based on AIC, BIC, CAIC, BAYES, BMSC-BAYES,

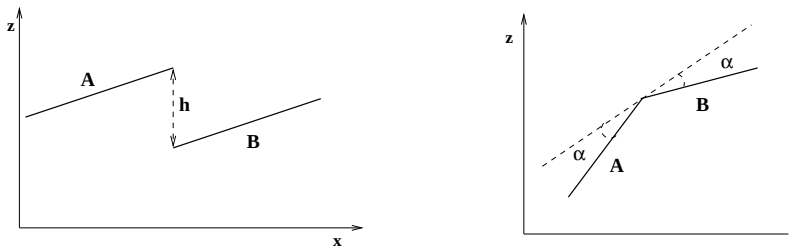
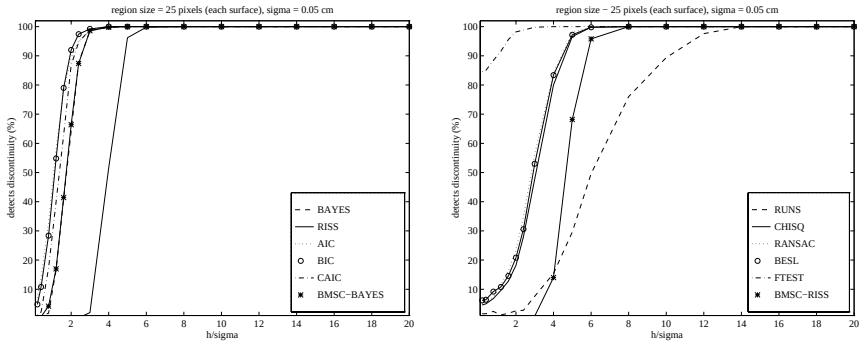


Figure 7: Shows step and crease discontinuity parameters.



(a) (b)

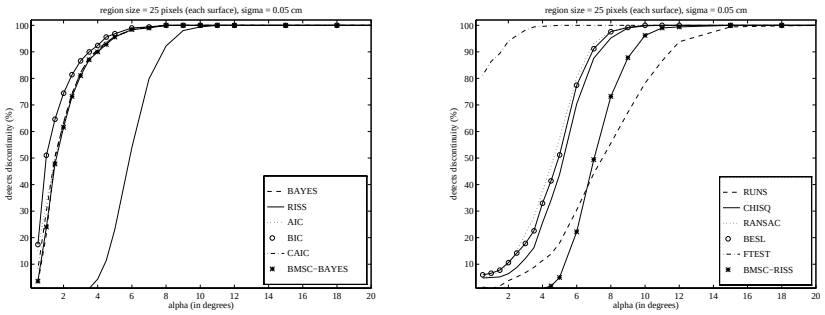
Figure 8: Performance of merging rules at step discontinuities.

and FTEST perform extremely well, detecting discontinuities with 98% success at $h = 3\sigma$ and 100% success at $h \geq 4\sigma$. Thus, the step discontinuity in Figure 1 (generated with $h = 4\sigma$ and same region size) can be detected by these merging rules. In contrast, for 100% success, RISS, CHI, RANSAC, and BESL need $h = 6\sigma$, BMSC-RISS needs $h = 8\sigma$, and RUNS needs $h = 12\sigma$.

For crease discontinuities, we generate data from the following two surfaces:

$$A : z = 100 + x \tan(\frac{\pi}{4} + \alpha), \quad B : z = 100 + x \tan(\frac{\pi}{4} - \alpha).$$

Thus, both A and B make an angle of α with the line $z = 100 + x$. Figure 9 shows percentage success of merging rules in detecting a discontinuity at different values of α at a region size of 25 pixels and $\sigma = 0.05$. The results show the same performance trends as for the step discontinuity. The merging rule based on FTEST shows a 100% success at $\alpha = 4$ degrees, while merging rules based on AIC, BIC, CAIC, BAYES, BMSC-BAYES show 98% success at $\alpha = 6$ degrees and a 100% success at $\alpha = 8$ degrees. Among other merging rules, CHI, RANSAC, and BESL show a 100% success at $\alpha = 10$ degrees, RISS at $\alpha = 11$ degrees, BMSC-RISS at $\alpha = 12$ degrees, and RUNS at $\alpha = 15$ degrees. Thus, the crease discontinuity in Figure 1 (generated using $\alpha = 10$



(a) (b)

Figure 9: Performance of merging rules at crease discontinuities with changing α.

criteria	min. α (in degrees)			criteria	min. α (in degrees)		
	σ=0.02	$\sigma = 0.05$	$\sigma = 0.1$		σ=0.02	$\sigma = 0.05$	$\sigma = 0.1$
BAYES	3	8	15	RUNS	9	15	27
RISS	4.5	11	18	CHI	4.5	10	21
AIC	3.5	8	15	RANSAC	4	10	18
BIC	3	8	15	BESL	4	10	18
CAIC	3.5	8	15	FTEST	2	4	10
BMSC-BAYES	3	8	15	BMSC-RISS	6	12	24

Table 4: Performance of merging rules at crease discontinuities with changing σ. Table shows the minimum α required by merging rules to correctly detect a crease discontinuity with 100% success.

degrees, same region size and σ) can be detected by most of the merging rules. The minimum α for 100% success increases with σ. Table 4 shows these α values at different σ.

For artificial (non-existent) discontinuities, data are generated for surfaces A and B from the line $z = 100 + x$. Table 5 compares performance of different merging rules at a region size of 25 pixels per surface. The results show that merging rule based on RISS, BAYES, BMSC-RISS, BMSC-BAYES perform the best, followed by RUNS, CAIC, CHI. Merging rules based on BIC, BESL, and RANSAC show a modest success rate of 91%. Surprisingly, the merging rule based on FTEST, which performed well when detecting discontinuities, only merges 15.4% of the artificial discontinuities, suggesting a strong bias in favor of preserving discontinuities. AIC shows only a 76.6% success. This is because, although AIC merges artificial discontinuities, it merges them to higher order surfaces.

Although not shown here, all merging rules show a gradual improvement in performance with increase in region size. For example, for step discontinuities, AIC, BIC, CAIC, BAYES, BMSC-BAYES show 100% success at $h = 2.4\sigma$ when the region size is 55 pixels (compare

rule	% success	rule	% success
BAYES	99.4	RUNS	96.6
RISS	100.0	CHI	94.4
AIC	76.6	RANSAC	90.8
BIC	91.4	BESL	91.4
CAIC	96.0	FTEST	15.4
BMSC-BAYES	98.8	BMSC-RISS	100

Table 5: Percentage success in merging artificial discontinuities to fit from correct model.

Overall Performance	Merging rules based on
Good	BAYES, CAIC, BMSC-BAYES
Average	RISS, BIC, BMSC-RISS, RUNS, CHI, RANSAC, BESL
Poor	AIC, FTEST

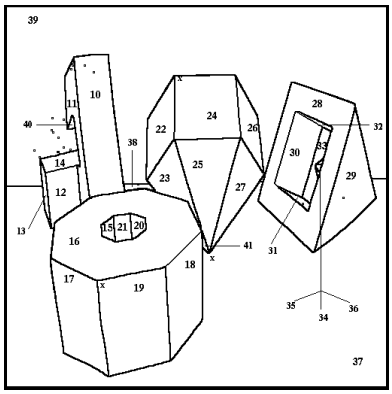
Table 6: Overall performance of merging rules using data with Gaussian errors.

this with 100% success at $h = 4\sigma$ at region size of 25 pixels). Similarly, for crease discontinuities, CHI, RANSAC, and BESL show 100% success at $\alpha = 5$ degrees when region size is 41 pixels (compare this with 100% success at $\alpha = 10$ degrees at region size of 25 pixels). This improvement is less pronounced for artificial discontinuities. The relative performance of different criteria, however, remains almost the same.

To summarize, a number of merging rules perform extremely well even at small step sizes ($h = 3\sigma$) and crease angles ($\alpha = 8$ degrees at $\sigma = 0.05$ cm), even for moderate region sizes (25 pixels). However, among these FTEST and AIC do not perform well for artificial discontinuities. While FTEST is biased towards preserving discontinuites, AIC merges to higher order surfaces. RUNS, CHI, RANSAC, and BESL also show a moderate bias towards merging surfaces at small region sizes. This is expected because these merging rules do not use any information from the old fits, but only look at finding a possible model for the combined data set. BMSC-BAYES shows consistently good results and gives a useful merging rule when sensor error models are unavailable or unreliable. Table 6 gives a qualitative summary of relative performances.

7 Results using real data

This section compares different model selection criteria and merging rules using one of the Perceptron test data sets from the University of South Florida's Segmentation Comparison Project [24] (see Figure 10(a)). This data set, consisting of planar surfaces, is particularly suitable because it provides ground-truth segmentation. Besides, model selection criteria and merging rules can be



(a)

criteria	large segments identified **incorrectly**
AIC	10, 16, 18, 25, 27, 28, 29, 33
CAIC	16, 28, 29, 33
BIC	10, 16, 28, 29, 33
BAYES	16, 29
RISS	16, 20, 21, 29, 33
RUNS	18, 19, 24

(b)

Figure 10: (a) Labels the segments in the image. For small segments, labels are marked outside the segment. (b) Model selection results for large segments.

tested in the presence of small-scale random noise, outliers, and potentially other kinds of unmodeled errors. For model selection, we test each ground-truth segment, as well as regions of different sizes within certain segments. Similarly, for merging, we test each ground-truth segment with its adjacent segments, and also test adjacent regions within certain segments. As mentioned before, we assume errors are t-distributed (following [9], $f = 1.5$), and σ is unknown. As a result, criteria that require knowledge of σ *a priori* or assume a Gaussian error distribution cannot be used. Also, the current versions of our bootstrap based criteria cannot be used in the presence of outliers. Thus, only AIC, BIC, CAIC, BAYES, RISS, and RUNS, and merging rules based on them are compared in this section.

7.1 Model selection

This section compares model selection criteria on both large and small segments. For most large segments, the different criteria correctly select a planar model. Figure 10(b) gives the large segments identified incorrectly by different criteria. BAYES performs the best followed by RUNS, CAIC, BIC, and RISS. As before, AIC continues to perform poorly. Table 7(a) gives the corresponding results for small segments. None of the criteria work well on the smaller segments in the scene. Of the nine small segments four are correctly identified by BAYES, three by RUNS, two by AIC, and one each by BIC and CAIC. RISS could not identify any of the small segments.

Another experiment tests the different criteria on square regions of progressively increasing sizes, starting from the pixels marked 'x', in segments 19, 24, and 37 (see Figure 10(a)). All criteria have problems when region size is small, and show improved performance as the region size gets larger. Table 7 shows the minimum region size required by each criteria in order to

criteria	small segments identified **incorrectly**
AIC	13, 31, 32, 34, 36, 40, 41
CAIC	13, 31, 32, 34, 35, 36, 40, 41
BIC	13, 31, 32, 34, 35, 36, 40, 41
BAYES	13, 34, 35, 40, 41
RISS	13, 31, 32, 34, 35, 36, 38, 40, 41
RUNS	34, 35, 36, 38, 40, 41

(a)

criteria	19	24	37
AIC	14 x 14	22 x 22	14 x 14
CAIC	38 x 38	22 x 22	14 x 14
BIC	38 x 38	22 x 22	14 x 14
BAYES	38 x 38	22 x 22	14 x 14
RISS	70 x 70	26 x 26	22 x 22
RUNS	10 x 10	14 x 14	10 x 10

(b)

Table 7: (a) Model selection results for small segments. Note that these segments are not labeled in Figure 10(a). See the caption of Figure 10(a) for a description of these segments. (b) Model selection in small regions in segments 19, 24, and 37. Region size in pixels.

select the correct model. The results show that RUNS works well for relatively small region size. AIC follows next, although it quickly starts selecting quadratic and cubic models as region size increases. BAYES, BIC, and CAIC show average performance, while RISS requires relatively large region sizes for selecting the correct model.

Overall, the behavior of model selection criteria on real data is similar to the simulation results. All criteria have problems at small region sizes and perform well as segments get larger. BAYES and RUNS perform the best, followed by CAIC and BIC. AIC and RISS perform poorly. Note the improved relative performance of RUNS in this section (compared to simulation results). This is expected because RUNS does not assume any noise properties of the data, and is thus less affected by inaccuracies in the error model. See Table 8 for a qualitative summary of relative performance.

7.2 Merging surfaces

The merging rules perform well on almost all merges between the large segments labeled in Figure 10. Of these, the information theoretic criteria only merge segments 11 and 14. Note that all segments inside the "nut" (segments 15, 21, and 20) are preserved. RUNS, on the other hand merges segment pairs (15, 16), (20, 21), and (22, 23). Among merges involving small segments, except for merging segments 13 with 14 and 32 with 28, all rules preserved discontinuities involving segments 13, 31, and 32. Only BAYES and RUNS preserved segments 13 and 14. Finally, all rules merged extremely small segments in the scene (segments 34, 35, 36, 40, 41) to adjacent surfaces.

Another experiment tests merging rules on adjacent square regions of progressively increasing sizes, starting from the pixel marked 'x', in segments 19, 24, and 37. At small region sizes,

Overall Performance	model selection	merging rules
Good	BAYES, RUNS	BAYES, BIC, CAIC, RUNS
Average	BIC, CAIC	RISS
Poor	AIC, RISS	AIC

Table 8: Overall performance for model selection and surface merging on Perceptron data.

although all rules merged the two surfaces, they did not merge them to the correct model. Other than one or two exceptions, the regions are correctly merged by the criteria when the combined region size reaches those shown in Table 7(b). AIC, however, merges these regions to a higher order model as region size increases.

Overall, most merging rules work well with moderate to large segment sizes, and have problems with small region sizes. Among these, BAYES and RUNS perform marginally better than others. Based on the above results, Table 8 gives a qualitative performance summary of merging rules.

8 Discussion

The results show that although some model selection criteria and merging rules definitely perform better than others, a moderate region size is crucial to the performance of all techniques. Unfortunately, there is no good way of quantifying small, moderate, and large. As rough indicators, a moderate region size is 25 pixels for the simulated data and (25 x 25) pixels for the Perceptron data. Although noise level is also important, a moderate region size is the dominating factor in the performance of all existing and newly introduced rules and criteria. Note that region size does not necessarily correspond to the number of pixels or data points, but the physical extent (say in cm) of the region. For example, simulations at a pixel size of 0.0032 cm (as opposed to 0.0016 cm in Section 6) show substantially better results.

Some of the techniques newly developed in this paper (for example, BAYES-BMSC) and those adapted here from the statistics literature (for example, BAYES) consistently show good performance. The information theoretic merging rules formulated in this paper perform well even at relatively small step sizes ($h = 3\sigma$) and crease discontinuities, and consistently merge artificial discontinuities. Unfortunately, none of the model selection criteria and new merging rules work as well as desired. Based on our results, we make the following recommendations when choosing among them.

- When the noise distribution of the data is known or can be closely approximated, BAYES

is a good choice for model selection and surface merging. Looking at the qualitative summaries in tables 3, 6, 8, BAYES shows good performance in all cases. Since BAYES requires calculating $|\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m)|$, for time-sensitive applications CAIC is a good alternative.

- When noise distribution is not known or cannot be closely approximated, BMSC-BAYES introduced in this paper is a good choice for data with independent, small-scale errors, and RUNS is a good choice for data with small-scale errors and outliers. Although, both techniques are computationally expensive, they are easily parellelizable.

- AIC and RISS should in general be avoided.

Finally, we make two additional comments when using these criteria in reconstruction algorithms.

- Merging and model selection should be avoided at small region sizes. It is better to fit planar patches to small windows and small seed regions, and use model selection and surface merging only on moderate to large region sizes.

- An orthogonal basis should be used for parameter estimation in model selection problems. This prevents numerical instabilities, and saves computation when using least-squares estimation.

9 Summary and Conclusion

We studied the problem of model selection and surface merging in surface reconstruction. We introduced new model selection criteria based on bootstrap data, and formulated new rules for surface merging using model selecting criteria. The new bootstrap criteria BMSC-BAYES consistently provides good results and may be useful when sensor error models are unavailable or unreliable; it still must be extended to handle outliers, however. The new merging rules developed here are free from heuristics and user-defined thresholds, and they perform extremely well even at small step and crease discontinuities. Finally, we extended some of these model selection criteria for use in the presence of outliers. Overall, the new criteria and merging rules can be used to give improved performance in surface reconstruction algorithms.

We tested the performance of different model selection criteria and merging rules on real and simulated data. This comparison study should be useful in choosing between different criteria given the data and given accuracy and efficiency requirements of the reconstruction application. Out of the criteria studied in this paper, this choice lies between BAYES, BMSC-BAYES, CAIC, and RUNS. Unfortunately, even the best test conditions, none of the model selection criteria and

merging rules work well at small region sizes. While this paper characterizes the effectiveness of different criteria and new merging rules improve upon previous results, it demonstrates the need for even better solutions to these problems.

References

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

[2] F. Arman and J. K. Aggarwal. Model-based object recognition in dense-range images - a review. *ACM Computing Surveys*, 25(1):5–43, March 1993.

[3] R. H. Bartels and J. J. Jezioranski. Least-squares fitting using orthogonal multinomials. *ACM Transactions on Mathematical Software*, 11(3):201–217, Sept. 1985.

[4] P. J. Besl. *Surfaces in Range Image Understanding*. Springer-Verlag, 1988.

[5] P. J. Besl, J. B. Birch, and L. T. Watson. Robust window operators. In *ICCV*, pages 591–600, 1988.

[6] P. J. Besl and R. C. Jain. Segmentation through variable-order surface fitting. *IEEE PAMI*, 10:167–192, 1988.

[7] R. M. Bolle and D. B. Cooper. Bayesian recognition of local 3-D shape by approximating image intensity functions with quadric polynomials. *IEEE PAMI*, 6(4):418–429, 1984.

[8] R. C. Bolles and M. A. Fischler. A RANSAC-based approach to model fitting and its applications to finding cylinders in range data. In *IJCAI*, pages 637–643, 1981.

[9] K. L. Boyer, M. J. Mirza, and G. Ganguly. The robust sequential estimator: A general approach and its application to surface organization in range data. *IEEE PAMI*, 16(10):987–1001, October 1994.

[10] H. Bozdogan. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370, 1987.

[11] K. A. Brownlee. *Statistical Theory and Methodology in Science and Engineering*. John Wiley and Sons, Inc., 1960.

[12] J. Cabrera and P. Meer. Unbiased estimation of ellipses by bootstrapping. *IEEE PAMI*, 18(7):752–756, 1996.

[13] F. S. Cohen and R. D. Rimey. A maximum likelihood approach to segmenting range data. In *IEEE Conference on Robotics and Automation*, pages 1696–1701, 1988.

[14] F. S. Cohen and J.-Y. Wang. Modeling image curves using 3-D object curve models — a path to 3-D recognition and shape estimation from image contours. *IEEE PAMI*, 16(1):1–12, Jan 1994.

[15] B. Curless and M. Levoy. Better optical triangulation through spacetime analysis. In *ICCV*, pages 987–993, Boston, MA, 1995.

[16] T. Darrell and A. Pentland. Cooperative robust estimation using layers of support. *IEEE PAMI*, 17(5):474–487, 1995.

[17] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley Publications, 1973.

[18] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[19] A. W. Fitzgibbon and R. B. Fisher. Lack-of-fit detection using the run-distribution test. In *European Conference on Computer Vision*, pages 173–178, Stockholm, 1994.

[20] C. Goodall. M-estimators of location: an outline of the theory. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors, *Understanding Robust and Exploratory Data Analysis*, chapter 11. John Wiley and Sons, 1983.

[21] F. Gustafsson and H. Hjalmarsson. Twenty-one ML estimators for model selection. *Automatica*, 31:1377–1392, October 1995.

[22] R. Hoffman and A. Jain. Segmentation and classification of range images. *IEEE PAMI*, 9:608–620, 1987.

[23] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Commun. Statist.-Theor. Meth.*, A6:813–827, 1977.

[24] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE PAMI*, 18:673–689, July 1996.

[25] E. T. Jaynes. *Probability Theory - the Logic of Science*. Physics, Washington University, St. Louis, MO 63130, USA, http://omega.albany.edu:8008/JaynesBook.html, 1994.

[26] S. M. LaValle and S. A. Hutchinson. A Bayesian segmentation methodology for parametric image models. *IEEE PAMI*, 17(2):211–217, Feb 1995.

[27] Y. G. Leclerc. Constructing simple stable descriptions for image partitioning. *IJCV*, 3:73–102, 1989.

[28] A. Leonardis. *Image Analysis Using Parametric Models : Model-Recovery and Model-Selection Paradigm*. PhD thesis, University of Ljubljana, 1993.

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

[30] M. Li. Minimum description length based 2D shape description. In *ICCV*, pages 512–517, 1993.

[31] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[32] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim. Robust regression methods for computer vision: A review. *IJCV*, 6:59–70, 1991.

[33] J. V. Miller and C. V. Stewart. MUSE: Robust surface fitting using unbiased scale estimates. In *CVPR*, pages 300–306, 1996.

[34] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:468–471, 1978.

[35] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

[36] B. Sabata, F. Arman, and J. K. Aggarwal. Segmentation of 3D range images using pyramidal data structures. *CVGIP:IU*, 57:373–387, 1993.

[37] K. Sato and S. Inokuchi. Range-imaging system utilizing nematic liquid crystal mask. In *ICCV*, pages 657–661, 1987.

[38] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

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

[40] C. V. Stewart, R. Y. Flatland, and K. Bubna. Geometric constraints and stereo disparity computation. *IJCV*, 20(3):143–168, 1996.

[41] J. Subrahmonia, D. B. Cooper, and D. Keren. Practical reliable Bayesian recognition of 2D and 3D objects using implicit polynomials and algebraic invariants. *IEEE PAMI*, 18(5):505–519, May 1996.

[42] F. S. Swed and C. Eisenhart. Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics*, 14:66–87, 1943.

[43] G. Taubin. Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range segmentation. *IEEE PAMI*, 13(11):1115–1138, 1991.

[44] R. Taylor, M. Savini, and A. Reeves. Fast Segmentation of Range Imagery Into Planar Regions. *CVGIP*, 45:42–60, 1989.

[45] S. Weisberg. *Applied Linear Regression*. John Wiley and Sons, 1985.

[46] P. Whaite and F. P. Ferrie. Active exploration: knowing when we're wrong. In *ICCV*, pages 41–48, 1993.

[47] S. C. Zhu and A. Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE PAMI*, 18(9):884–900, 1996.