Abstract

The problem of model selection is relevant to many areas of computer vision. 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. More importantly, suitable extensions to these criteria must be made to solve problems unique to computer vision. Using the problem of surface reconstruction as our context, we analyze existing criteria using simulations and sensor 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 a Bayesian criteria and its bootstrapped variant perform the best, although for time-sensitive applications, a variant of the Akaike criterion may be a better choice. Unfortunately, none of the criteria work reliably for small region sizes, implying that model selection and surface merging should be avoided unless the region size is sufficiently large.

1 Introduction

Determining the correct representation to describe a data set is an important issue in computer vision, arising in several domains such as segmentation, surface reconstruction, 3D modeling, object recognition, reverse engineering, inspection, and tracking. In each (a) a choice must often be made between representing a given image region with a single functional description or partitioning it and describing each sub-region with a separate function, (b) the appropriate function (model) must be chosen for each data set, and (c) the parameters of the chosen model must be accurately estimated. In the absence of any prior knowledge, most approaches solve this problem by first selecting the model and estimating its parameters in relatively small image regions (fig. 1(a)), and later attempting to merge these regions into larger image regions (fig. 1(b)). While the problem of estimating the model parameters is well studied [6, 7, 25, 26, 32], the associated issues of model selection and merging have received much less attention in computer vision. Yet without good solutions to these problems, the estimated parameters have little meaning.

The relative strengths of model selection techniques and merging techniques used in vision algorithms are not yet well understood. Some model selection criteria in vision rely on heuristics and user-defined thresholds [4, 6, 29, 34], others, especially recent ones, are applications of information theoretic criteria [5, 7, 12, 21, 22, 23, 33, 37, 38]. Most of these criteria 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. Merging techniques, with the exception of [20], are generally based on empirical heuristics [4, 11, 35, 29]. Further, in an attempt to avoid the model selection problem, many merging techniques only join fits to the same model [4, 11, 20, 35], potentially limiting the effectiveness of merging. Hence, mathematical criteria to merge regions and to simultaneously decide the correct model for a merged region must be formulated. Finally, both model selection and merging criteria in vision must tolerate outliers [7, 12] and unknown noise distributions.

One computer vision problem where model selection
and merging techniques are crucial for 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 involves estimating initial surface patches (using grid techniques [11, 35], clustering methods [15, 29], or by region growing [4, 7, 22, 34]), and later pruning redundant surface patches [7, 22], or merging artificial surface boundaries [4, 11, 29, 35]. In the absence of a priori information, model selection forms an important part of each 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. When pruning redundant fits a model selection criterion may be used explicitly [7] or combined with greedy search techniques [22]. When merging adjacent surfaces, a criterion must be used to determine if the data should be represented by a single fit or by two or more different fits. Observe how the problem of model selection is implicit in the problem of surface merging.

Surface reconstruction, therefore, provides a good context for studying the model selection and merging problems in computer vision. Using this problem as our context, we study the characteristics of different model selection criteria, 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. The experiments on simulated and sensor data 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 criteria and merging criteria for different types of data and applications. We study both information-theoretic model selection criteria, as well as criteria based on Chi-square test [37], F-test [36, page 96], and runs test [3, 9], and formulate merging rules based on them. However, this paper only presents the information-theoretic model selection and merging criteria, and simply summarizes the performance of the rest. For these criteria and detailed results, refer [10].

2 Definitions

Range image: A range image is characterized by a point \( p_i = [x_i, z_i]^T \) at any pixel \( i \) in the image. For our simulations, \( x_i \) will simply be a scalar \( x_i \), and for sensor range images \( x_i = [x_i, y_i]^T \). We call the former 2D range images and the latter 3D range images.

Candidate models and parameter estimation: We fit orthogonal polynomials [2, 3] to the data. This gives well-conditioned matrices, and is efficient because the fit to higher order models builds on fits to lower order models. The second advantage, however, is lost when using robust estimation techniques.

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.

Model selection techniques fit each candidate model to the data, and choose a model based on various measures of fit accuracy and model complexity. Thus, parameter estimation forms an important part of model selection. Let \( D = \{ p_1, \ldots, p_n \} \) be a set of \( n \) data points. Consider models of the form

\[
Z = X_m \theta_m + e, \tag{1}
\]

where \( Z \) contains the \((n \times 1)\) depth values, \( X_m \) contains \((n \times d_m)\) orthonormal polynomials, \( \theta_m \) is the \((d_m \times 1)\) parameter vector, and \( e = [e_1, e_2, \ldots, e_n]^T \) is a vector of unobserved, but independent random variables, with scale \( \sigma \). Note that \( \sigma \) may or may not be known a priori. Information-theoretic criteria use the loglikelihood of estimated parameters for model selection, hence, maximum likelihood estimators (MLEs) must be used for parameter estimation. We use ordinary least-squares for data with Gaussian errors, and following [7], we use iteratively reweighted least squares (IRLS) [16] with an M-estimator based on \( t \)-distribution for data with outliers. In the latter case, IRLS is initialized using least median of squares (LMS) [25]. We denote the likelihood for model \( m \) by \( L(\theta_m) \) and the MLE of \( \theta_m \) by \( \hat{\theta}_m \). When \( \sigma \) is unknown, the likelihood is written as \( \log L(\theta_m, \sigma) \), and \( \hat{\sigma}_m \) is the MLE of \( \sigma \) for model \( m \). See [10] for details.

3 Model selection

This section summarizes model selection criteria already used in reconstruction algorithms and introduces promising criteria from the statistics literature.

Most Information-theoretic criteria are based on Kullback-Leibler (K-L) distances, Bayesian probabilities, and Minimum Description Length (MDL). The criteria based on K-L distances select the model minimizing

\[
d(\theta_m, \theta_*) = E.[-2 \log L(\theta_m)]|_{\theta_m = \theta_*} \tag{2}
\]

where \( \theta_\ast \) represents the parameters for the “true” or generating model and \( E \) denotes the expectation under the generating model. Note that evaluating (2) is not possible, since it requires knowledge of \( \theta_\ast \). However, \( -2\log L(\hat{\theta}_m) \) is used as a (biased) estimator of (2). Different criteria based on K-L distance result from different bias adjustments to \(-2\log L(\hat{\theta}_m)\). 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
The idea of bootstrap is simple. Consider the regression model given by (1), and let $P$ be the unknown distribution of the elements in $e = [e_1 \ldots e_n]^T$. Let $\hat{\theta}_m$ be the least-square estimate of $\theta_m$, let $\hat{e}_m$ be the corresponding residuals, and let $\hat{Z}_m$ be the corresponding estimates of the uncorrupted $Z$. The residuals, $\hat{e}_m = [\hat{e}_m, \ldots \hat{e}_m]^T$, can be used to generate an empirical estimate, $P_m$. $P_m$ is defined to be the discrete distribution that puts probability $1/n$ on each value $\hat{e}_m, \ldots, \hat{e}_m$. The plug-in bootstrap principle [14, chapter 4] substitutes $P$ with $P_m$ and this is used to generate $R$ “bootstrap” error vectors, $e_{1m}, \ldots, e_{Rm}$. Sampling from $P_m$ is the same as sampling from the set $\{e_1, \ldots, e_n\}$ with replacement. Each $e_{im}$ is added to $Z_m$ to generate a bootstrap set of $Z$ values, $Z_{im}$. Each “bootstrap” data set $Z_{im}$ can now be used to generate bootstrap least-square estimates $\hat{\theta}_{im}$, creating $R$ such estimates for generating statistics on $\theta_m$.

In this paper, we derive a bootstrap version of the BAYES criteria given by $P(D|m, I)$ in table 1 (see [10] for criteria based on RISS). Like any other criteria, it is a penalized likelihood, balancing between accuracy and complexity of the model given the data. In this criteria, $L(\bar{\theta}_m)$ measures accuracy and $H(\hat{\theta}_m)$ measures complexity or model stability. First consider the accuracy term. When the error distribution is unknown, we replace $\log L(\bar{\theta}_m)$ with the normalized residual sum of squares, $-RSS_m / \sigma^2$, using least-squares estimation. More sophisticated estimators are unnecessary because our weak assumptions on sensor noise are sufficient to yield unbiased, minimum variance estimates with least-squares [24, page 172]. The bootstrap estimate of $\sigma$, $\sigma^*$, is calculated by finding the average standard deviation of $Z_{1m}, \ldots, Z_{Rm}$. Unfortunately, this gives four different $\sigma^*$'s corresponding to the four models in $M$. However, our experiments show that the $\sigma^*$ 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 $\sigma^1$. This indicates that $\sigma_{m3}$, the bootstrap estimate for the cubic model, can be used as an estimate of $\sigma^*$. To obtain a distribution-free measure of $H(\bar{\theta}_m)$, observe that $H(\bar{\theta}_m) \approx -[V(\bar{\theta}_m)]^{-1}$, the covariance matrix of $\bar{\theta}_m$ [18, chapter 24]. The bootstrap estimate of $V(\bar{\theta}_m)$, $V^*(\bar{\theta}_m)$ can be calculated from $\hat{\theta}_1^m, \ldots, \hat{\theta}_R^m$. We then obtain a bootstrapped, Bayesian model selection criterion (BMSC-BAYES) by taking the natural logarithm of $P(D|m, I)$, replacing $\log L(\bar{\theta}_m)$ with $-RSS_m / \sigma_{m3}^2$ and $H(\bar{\theta}_m)$ with $-[V^*(\bar{\theta}_m)]^{-1}$:

$$BMSC\text{-BAYES}_m = \frac{d_m}{2} \log 2\pi - \frac{RSS_m}{\sigma_{m3}^2} + \frac{1}{2} \log |V^*(\bar{\theta}_m)|. \tag{3}$$

5 New rules for surface merging

This section extends the model selection framework to develop new rules for merging surface patches to a single

---

Table 1: Table shows different model selection criteria. While the first expressions corresponding to each criteria is used when $\sigma$ is known a priori, the second expression is used when $\sigma$ is unknown. For unknown $\sigma$, BAYES maximizes $P(D|m, G)$ for a Gaussian distribution, and $P(D|m, t)$ for the t-distribution (see [10]). In these expressions, $RSS_m$ denotes the residual sum of squares for model $m$, $H(\bar{\theta})$ stands for the Hessian of the log-likelihood, $V_{dm}$ is the volume of the $d_m$-dimensional unit hypersphere [13, page 24], and $\log_2 s(t) = \log_2 t + \log_2 \log_2 t + \ldots$, including only its positive terms.

Probability is denoted by $P(D|m, I)$. Finally, model selection criteria based on the MDL principle choose a model that minimizes the number of bits, $I_{enm}$, required to express the observed data. Table 1 shows the various criteria studied in this paper. While AIC, and CAIC are based on the K-L distances, BAYES and BIC are derived from Bayes rule using non-informative priors. RISS on the other hand is based on MDL. The MDL criteria in [27] is exactly the same as BIC. Interestingly, the MDL criteria included in a quadratic optimization function in [22] can be shown to be similar in form to AIC [10].
surface description. We assume that small surface patches have already been estimated using different approaches summarized in Section 1, and these surface patches do not undersegment the scene, i.e., they do not bridge discontinuities. For the following discussion, we only consider the core problem of merging two surface patches.

To define the problem precisely, suppose two surfaces $A$ and $B$ are fit to noisy data sets $D_A$ and $D_B$ where $D_A \cap D_B = \emptyset$. The issue is to determine whether $D_A$ and $D_B$ are measurements from the same or different underlying surfaces. 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 $C_0, \ldots, C_3$, corresponding to models $m_0, \ldots, m_3$, be fits to the data set $D_C = D_A \cup D_B$. 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 $D_C$.

For surface merging, we extend the model selection criteria of section 3 to select models $m_A$ and $m_B$ together or any one of models $m_0, \ldots, m_3$ for the data set $D_C = D_A \cup D_B$. To do this, measures of probabilities, K-L distances, and description lengths must be formulated for $m_A$ and $m_B$ combined. Since $D_A$ and $D_B$ are disjoint, their joint likelihood to fits from $m_A$ and $m_B$ are $L(\theta_{m_A})L(\theta_{m_B})$. The K-L distance of this likelihood under the generating model is then $E[-2 \log L(\theta_{m_A})L(\theta_{m_B})]$. Evaluating at maximum likelihood estimates this reduces to

$$d(\theta_A, \theta_B) + d(\theta_{m_A}, \theta_{m_B}).$$

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 $\text{len}_{m_A, m_B} = \text{len}_{m_A} + \text{len}_{m_B}$. Based on this, merging decisions for K-L distances, Bayesian probabilities, and MDLs may be represented as

$$\min\{d(\theta_A, \theta_B), d(\theta_{m_A}, \theta_{m_B}), \ldots\},$$

$$\max\{P(D_A|m_A, I) P(D_B|m_B, I), P(D|m_B, I), \ldots\},$$

$$\min\{\text{len}_{m_A} + \text{len}_{m_B}, \text{len}_{m_1}, \ldots\}.$$ 

respectively. Replacing the model selection criteria from table 1 in the corresponding decision functions above gives merging rules based on AIC\(^2\), CAIC, BAYES\(^3\), BIC, and RISS. For formulating merging rule based on BMSC-BAYES, note that it is based on the position of the Bayesian probability, $P(D|m, I)$. As such, the corresponding merging rules is similar to the Bayesian rule, and is given by

$$\max\{\text{BMSC-BAYES}_A + \text{BMSC-BAYES}_B, \text{BMSC-BAYES}_m, \ldots\}$$

\(^2\)It can be shown that the optimization function used in [22] can be used for merging surfaces, and is similar to the merging rule based on AIC [10].

\(^3\)In surface reconstruction, a similar Bayesian merging approach has been used in [20]. However, this approach only merges surfaces corresponding to the same model. Besides, it also restricts the parameter space such that $|\theta_{m_0}| = 1$. As such, this work can be considered as a special case of ours.

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.

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 [30]. The criteria assume Gaussian errors and known $\sigma$ (note that BMSC-BAYES makes no assumption regarding noise distributions or $\sigma$).

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 $\sigma$, and by varying $a_1$ for linear model, and $a_2$ for quadratic model. Our sensor has a $\sigma$ of about 0.02 cm at a depth of 100 cm. In our experiments we vary $\sigma$ from 0.02 to 0.1 cm. Results are based on 500 simulations, and for bootstrap criteria, the number of bootstrap replications, $R$, is set to 200 [14].

In the first 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 sizes symmetrically around the origin (see Figure 2) from 7 to 77 pixels and varies $\sigma$ from 0.02 cm to 0.1 cm. Figure 3(a) shows percentage success of different selection criteria for data from the linear model at $\sigma = 0.05$ cm. 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. The new bootstrap-based criteria, BMSC-BAYES also performs well, closely following BAYES. This performance is promising, given that it does not make any assumption regarding the noise distribution. 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 3(b) shows corresponding performance 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 $\sigma$. For example, for BAYES this size is 25 pixels at $\sigma = 0.02$ cm and 40 pixels at $\sigma = 0.1$ cm. The results show several differences from the linear case. First, with increasing $\sigma$, all criteria find it increasingly difficult to find the quadratic model at small region sizes. This is not surprising, given the difficulty in identifying a quadratic fit from the data in Figure 2(b). Second, RISS, which performs the best for linear models even for small regions, now perform poorly at small region sizes. This suggests a possible bias in RISS towards low order surfaces. But once again, at large region sizes BAYES, RISS, and CAIC have problems at small region sizes, their performance improves as region size increases. The new bootstrap-based criteria, BMSC-BAYES also performs well, closely following BAYES. This performance is promising, given that it does not make any assumption regarding the noise distribution. 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 3(b) shows corresponding performance 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 $\sigma$. For example, for BAYES this size is 25 pixels at $\sigma = 0.02$ cm and 40 pixels at $\sigma = 0.1$ cm. The results show several differences from the linear case. First, with increasing $\sigma$, all criteria find it increasingly difficult to find the quadratic model at small region sizes. This is not surprising, given the difficulty in identifying a quadratic fit from the data in Figure 2(b). Second, RISS, which performs the best for linear models even for small regions, now perform poorly at small region sizes. This suggests a possible bias in RISS towards low order surfaces. But once again, at large region sizes BAYES, RISS, and

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 4), 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 5(a) shows the percentage success of the merging rules in detecting a discontinuity at different values of $h/\sigma$ at a region size of 25 pixels per surface. The results show that merging rules based on AIC, BIC, CAIC, BAYES, and BMSC-BAYES perform extremely well, detecting discontinuities with 98% success at $h = 3\sigma$ and 100% success at $h \geq 4\sigma$. In contrast, for 100% success, RISS, needs $h = 6\sigma$.

For crease discontinuities, we generate data from the following two surfaces (figure 4(b)):

$$A : z = 100 + \pi \tan \left( \frac{\pi}{4} + \alpha \right), \quad B : z = 100 + \pi \tan \left( \frac{\pi}{4} - \alpha \right).$$
Figure 6: Labels the ground-truth segments in the image. For small segments, labels are marked outside the segment.

7 Results using sensor data

This section compares different criteria using one of the Perceptron test data sets from the USF Segmentation Comparison Project [17] (see Figure 6(a)). This allows testing different criteria in the presence of small-scale random noise and outliers. To test performance of model selection, we apply each criteria on the ground-truth segments provided with the image, as well as on regions of different sizes within certain segments. For testing performance of merging criteria, we attempt to merge each ground-truth segment with its adjacent segments. In another experiment, we divide a ground-truth segment into small regions and attempt to merge them using these criteria. As mentioned before, we assume errors are $t$-distributed (following [7], $f = 1.5$), and $\sigma$ is unknown. Note, that the current version of BMSC-BAYES cannot be used in the presence of outliers.

### Table 2: Overall performance of model selection and merging criteria using data with Gaussian errors.

<table>
<thead>
<tr>
<th>Model selection criteria</th>
<th>Model selection performances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>RISS, CAIC, BIC, BAYES</td>
</tr>
<tr>
<td>Average</td>
<td>RISS, CAIC, BMSC-BAYES</td>
</tr>
<tr>
<td>Poor</td>
<td>AIC</td>
</tr>
</tbody>
</table>

### Table 3: Model selection results for large and small segments.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Large segments identified incorrectly</th>
<th>Small segments identified correctly</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>10, 16, 18, 25, 27, 28, 29, 33</td>
<td>35, 38</td>
</tr>
<tr>
<td>CAIC</td>
<td>16, 28, 29, 33</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>10, 16, 28, 29, 33</td>
<td>38</td>
</tr>
<tr>
<td>BAYES</td>
<td>10, 29</td>
<td>31, 32, 36, 38</td>
</tr>
<tr>
<td>RISS</td>
<td>16, 20, 21, 29, 33</td>
<td></td>
</tr>
</tbody>
</table>
ues to perform poorly. The third column of table 3 gives the small segments identified correctly by different criteria. None of the criteria work well on the smaller segments in the scene. Of the nine small segments four are correctly identified by BAYES, 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 6). The results show that BAYES, BIC, and CAIC begin selecting the correct model at a region size of about 25 x 25, while RISS requires a relatively large region size of 40 x 40 pixels. AIC starts selecting quadratic and cubic models at region size increases.

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. See Table 4 (second column) 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 6. Only segments 11 and 14 are merged by all criteria. Note that all segments inside the “nut” (segments 15, 21, and 20) are preserved. Among merges involving small segments, all rules merge 32 with 28, and barring BAYES all rules merge 13 and 14. All other discontinuities involving segments 13, 31, and 32 are preserved. Finally, all rules merge extremely small segments in the scene (segments 34, 35, 36, 40, 41) to adjacent surfaces.

Another experiment tests merging artificial discontinuities on adjacent square regions of progressively increasing sizes, starting from the pixel marked ‘x’, in segments 19, 24, and 37. At small region sizes, 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 25 x 25 for BAYES, BIC, CAIC, and 40 x 40 for RISS. 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 them, BAYES performs marginally better than others. Based on the above results, table 4 (last column) gives a qualitative performance summary of merging rules.

### 8 Discussion and Conclusion

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 test data. Although noise level is also important, a moderate region size is the dominating factor in the performance of all existing and newly introduced 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.

Criteria based on Chi-square, F, and runs test, detailed in [10], also perform poorly at small region sizes. These criteria perform relatively worse than most information-theoretic model selection criteria, reaching a success rate of only 90% to 95% in the experiments described in section 6.1. The new merging rule based on F-test compares with BAYES in detecting discontinuities, but only merges 15.4% of the artificial discontinuities, suggesting a strong bias in favor of preserving discontinuities. The merging rules based on Chi-square and runs test also perform relatively worse than information-theoretic rules, showing 100% success for step discontinuities at $h = 6\sigma$ and for crease discontinuities at $\alpha = 10$ degrees, in the experiments described in section 6.2. Further, Chi-square and runs based merging rules show a bias towards merging surfaces at small region sizes. Overall, the information-theoretic criteria work better than criteria based on different statistical tests.

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 2 and 4, BAYES shows good performance in all cases. Since BAYES requires calculating $\|H(\theta_m)\|$, 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. Although this technique is computationally expensive, it is easily parallelizable.
- AIC and RISS should in general be avoided.
- Merging and model selection should be avoided at small region sizes.

#### Table 4: Overall performance for model selection and surface merging on Perceptron data.

<table>
<thead>
<tr>
<th>Performance</th>
<th>Model Selection</th>
<th>Merging Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>BAYES, BIC, CAIC</td>
<td>BAYES, BIC, CAIC</td>
</tr>
<tr>
<td>Average</td>
<td>BIC, CAIC</td>
<td>RISS</td>
</tr>
<tr>
<td>Poor</td>
<td>AIC, RISS</td>
<td>AIC</td>
</tr>
</tbody>
</table>
To conclude, 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. BAYES-BMSC, however, still must be extended to handle outliers. The merging rules formulated in this paper perform well even at relatively small step sizes ($h = 3\sigma$) and create discontinuities. Unfortunately none of the model selection criteria and merging rules work well at small region sizes. Thus, 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.

Acknowledgments

The authors would like to thank James Miller for helpful discussions and comments on earlier drafts of this paper. They would also like to acknowledge the financial support of the National Science Foundation under grant IRI-9408700.

References