7. ANALYSIS

The remainder of this dissertation endeavors to assemble and interpret the graphical and statistical information used in the test of the hypothesis put forth by this work. If a visual model is ultimately to be used to compute medical image quality measurements, what is necessary from the model results is that they parallel those of the human. The presumption when measuring observer performance to provide an indication of image quality is that the performance will vary over the range of potential parameter settings of image acquisition or processing. That observer performance makes a statement about the quality of the images; where accuracy is optimal indicates the highest quality image. A model for the observer can be more or less accurate than the human. As long as the model maintains a constant relationship across the potential parameter values, then the model provides an equally good indicator of quality (see Figure 7.1). The most important analyses then in this chapter test the hypothesis that the differences between the estimation errors for the model and the human are constant.

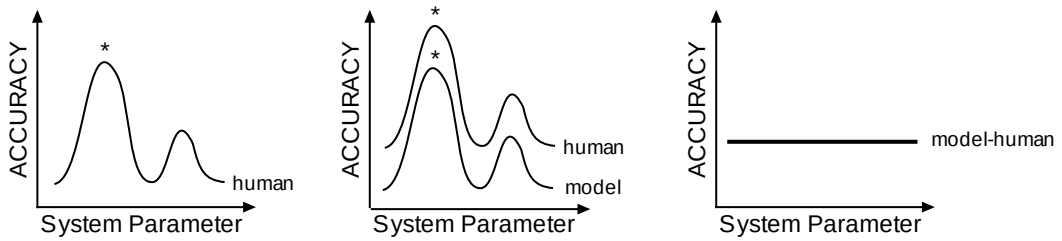


Figure 7.1. Parallel performance of model and human. Human accuracy (left) as a function of some parameter of the imaging system or processing method can be used to target an optimal setting of that parameter. A model that is known to have a parallel relationship to human performance (middle) can just as well be used to locate those parameters that are best for human use. In that case, the difference between the model and human is constant (right).

The data were all analyzed using one of several forms of a repeated measures linear models approach (see Section 2.4).[1] The analyses make a linear least squares (or alternatively a maximum likelihood) approximation to the data as a function of the independent variables in the design. The analysis models allow specification of those variables for which repeated measures were obtained. Indeed, in both of the experiments in this research, all independent variables (for example, angiographic blur or SHAHE contrast) were indicated as repeated, or within-subjects, as an observer contributed scores at each of the conditions within each variable. As mentioned in Section 2.4, when the levels of the independent variables are described categorically, standard ANOVA techniques may be utilized. When the independent variables exist on a continuous scale, the analysis must resort to regression techniques.

A multivariate approach to repeated measures ANOVA was used to analyze the portal imaging study data. The linear model in this case is

$$Y_i = X\beta + \varepsilon_i \qquad 7.1$$

Y is the dependent variable matrix; Y_i is a row vector of responses from the j conditions for i^{th} observer. X is a matrix that contains the levels of the factors that are studied in the experiment, and the β's are the "effect parameters" that are calculated to estimate the population's relationship of each condition in X to the dependent variable. The analysis assumes

1) homogeneity of variance, i.e., that each subject has the same covariance matrix,
2) the common covariance matrix is a multivariate normal distribution about the mean, and
3) the vector of responses is independent for each subject.

The probability that is reported in these analyses is the Geisser-Greenhouse (henceforth "G-G") corrected p-value, an adjusted significance test that corrects for deviations from the variance assumptions. Furthermore, when each of the analyses was done, the normality assumption was verified with univariate statistics at each experimental condition.

A regression approach was employed for most of the analyses for the angiography experiment. The original data analysis plan was a simple ANOVA using the orthogonal, categorical independent variables that

[1] The SAS System, SAS Institute, Inc., Cary, NC.

were used in the generation of the experiment images. When the design was changed to describe experimental conditions with the more principled perceptual measures of blur, the effective blur scale (EBS), and noise, the figure-to-noise ratio (FNR), the following analysis was required to accommodate those continuous, non-orthogonal independent variables describing the data.

In general, a linear regression equation describes a linear least-squares approximation of the dependent data as a function of one or more continuous independent variables. The variability in the data may be apportioned based on the regression into two components: variability predicted by the regression on the independent variable and residual variability not predicted by the regression. Those components of the regression allow the linear model hypothesis tests: the null hypothesis that a particular variable has no effect is rejected if the variability of the data decreases sufficiently when that extra term is added to the model. There are, as with categorical ANOVA, assumptions about the relationship that the regression describes. For the particular analyses of the angiography data, the necessary assumptions were that the samples at any particular value of the independent variable were assumed to have a mean lying along the regression line and have errors that were normally distributed. The independence assumption is corrected for by the repeated measures techniques used in the analysis.

The angiography study possessed repeated continuous factors, or covariates, EBS and FNR. Within each subject, the covariates are different in each of the repeated measures conditions. The appropriate analysis in this case is a repeated covariates regression model that employs a univariate approach. The analysis relies upon different parameter estimation and hypothesis testing techniques since the repeated covariates make it not possible to develop a closed form expression for the linear model. The algorithms in the SAS procedure "MIXED" were used to calculate restricted maximum likelihood (REML) estimates of the β's in the model. This analysis again assumes the equivalence of each observer's covariance matrix. However, with so many repeated conditions and so few observers, it was impossible to correct for the absence of this condition. Some attempts were made in the analysis of the angiography data to assure that these compound symmetry requirements were satisfied. Nonetheless, these regression analyses must be regarded as liberal, or quick to reject the null hypothesis. Therefore, where the repeated covariates regression techniques were used, the significance decisions were made at the 0.01 level.

The procedure for interpreting the results of this kind of a regression analysis is similar to that for a standard ANOVA: trends are discussed in terms of the main effects and interactions (see definitions below). Because of the non-orthogonality of the angiography design, however, the way that the regression must be performed is by successively adding terms into the linear model and determining whether each one in turn contributed significantly to a better representation of the data. The stenosis depth parameter resulted in the dominant trend in the data. Furthermore, the blur and noise conditions were of interest at any particular stenosis depth. Therefore the analysis adds the stenosis depth term into the model first. Subsequent significance tests involving the blur and noise parameters ask only whether these additional terms contribute anything further toward the linear model description of the data. The EBS and FNR factors may be confounded with stenosis depth: the stenosis depth parameter explicitly indicates the width of vessels, and the other two parameters are defined by formulas that explicitly incorporate width information as well. To the extent that there is variation in performance with respect to stenosis depth, the EBS and FNR parameter may reflect or more likely cause some of that variation as well. However, a significant effect of EBS, for instance, will only be determined when variation in the EBS value causes variation in the scores above and beyond that caused by the variation in depth. It is also because of these potential relationships between the parameters that the data were often analyzed separately at each depth condition.

It is important to define for each analysis the *outcome* measure, or "score," that was the value for the dependent variable at each condition. Section 3.2 discussed how the perception of shape is thought to be roughly zoom invariant: the absolute error with which a figure can be perceived is monotonically related to the figure's width. The estimation tasks in these experiments involved judgments about widths of objects that vary across trials. For the angiography experiment, observers and the model made estimates about simulated vessels with different stenosis depths. In the portal imaging study, the true treatment field clearance varied considerably on each trial. Thus the outcomes are expressed as *relative errors*, which divide absolute error by the magnitude of the quantity in question. Relative errors in this way make commensurate outcomes from experimental conditions that have different truths.

The original designs for both the angiography and portal experiments specified an outcome that was an absolute error. While the decision to analyze relative errors was made *post-hoc*, it is firmly justified on theoretical grounds and was never proposed as a means of intentionally improving the correspondence between the model and human data.

In the angiography experiment, where the truth was known, the outcome measure for human performance alone was human error relative to truth, or the difference between the percent stenosis estimate from the human and the true percent stenosis, all normalized by the width-proportional quantity 100 minus true percent stenosis. Similarly, the outcome reflecting the model's performance is the model's relative error, or its percent stenosis

minus the truth normalized in the same way. For the analysis that compared the performance of the model and the human, the outcome measure is a normalized difference between model and human percent stenosis.

In the portal imaging study, the primary outcome examined was the difference between the model and human distance in pixel units normalized by the true distance. The value for the truth was a rough estimate provided by the author. Further conclusions about the portal imaging experiment are suggested in Section 7.5 by assessing separately model and human accuracy with respect to this estimate of truth.

The analyses in this chapter thus all test whether an outcome measure that is a relative error does or does not vary significantly as a function of the variables in the study. The descriptions of the data in this chapter will often use the term accuracy, which is generally regarded to mean a positive quantity whose larger values indicate better performance. However, the actual data are reported as signed relative errors, where zero error is perfect accuracy. The magnitude, or absolute value, of the errors was never examined; the value, or "direction," of the difference was important in attempting to understand or draw conclusions about the cause of any discrepancies between the human or the model and truth, or more importantly between the model and the human.

In the analyses for both the angiography and portal imaging studies, the scores at each condition for the model and the observers were means of the relative errors across all of the backgrounds that served in that condition. In the analyses that compared model and human performance, mean model performance was subtracted from each observer's mean at each condition.

The analysis methods provide estimates of the significance of main effects and interactions for the independent variables in the design. A test for a main effect of a particular independent variable, A, combines scores across all other variables at each of the levels of A. What is produced is an indication of the probability that the variation in the single variable A caused significant differences in the combined scores at the different levels of A. A significant effect for the contrast variable in the portal imaging study would indicate that when combining scores across the other variables in the study the different contrast levels caused an overall variation in the outcome. Interactions study the relationships between variables (Figure 7.2). An interaction between two variables is said to exist when the linear relationship between the dependent variable and one of the independent variables is in turn dependent on the other independent variable. For example, an interaction would exist between the noise and blur variables in the angiography study if there were different response patterns as a function of the blur parameter at the different noise levels. The chosen significance level is used in deciding whether the probability, or "p" value, warrants rejecting the hypothesis of no difference, or no main effect, or no interaction.

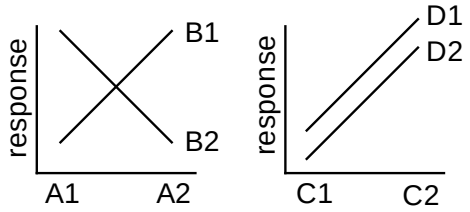


Figure 7.2. Interaction of variables. Variables A and B are said to interact: the relationship between the responses to independent variable B is dependent on the level of the other independent variable A. On the other hand, the pattern of responses at variable D does not change with variation in variable C.

Statistical indications of main effects and interactions for either human or model estimates or differences between them are presented in a "source table" (see for example Table 7.2). For an analysis performed on any particular one of these dependent variables, the table lists the independent variables along with an F statistic and rejection probability estimate for each. In the left column, the single (for a main effect) or multiple (for interactions) independent variables are listed in the "SOURCE" column. The "F" statistic is a summary measure that provides an indication of the increase in the error entailed by representing the data by the full as opposed to the restricted linear model. F is large when the null hypothesis ought to be rejected. The right column ("Pr > F") provides an estimate of the probability that the variation in the data due to the independent variable(s) in question occurred by chance alone. When "p" is less than the predetermined significance level, the null hypothesis is rejected. So for example, the first row in Table 7.2 provides an F value (26.45) and probability (< 0.0001) of rejecting the hypothesis that the outcome, in that case human relative error, did not vary significantly as a function of the "depth" parameter alone.

When the analysis studies an outcome that is a relative error between the model and human results, the hypothesis that is tested is that the difference is constant. For a model to successfully predict human behavior with respect to the independent variables studied, there must be no main effects for any of the variables: errors should not significantly differ from each other at the levels of the variables. Likewise, it would be necessary for the pattern of scores with respect to a single variable to remain virtually constant at the different levels of another

variable. Of course, in the absence of this perfect correspondence, it is informative to know at which of the variables the model did or did not predict sufficiently.

7.1 Human Performance for Stenosis Estimation

It was of interest to test whether human performance alone varied with respect to the parameters studied in the angiography experiment. While the ultimate test is whether the model results parallel those of the human, however variable the human performance was, it would be a more conclusive statement of the model to know that it did or did not predict human performance that was not simply constant. Furthermore, as it is this human data which is the standard against which the model was to be compared, it is important to first establish that the human behavior was for the most part understandable and believable before using it in a test of the model. The analysis in this section tested whether the outcome measure, human percent stenosis minus true percent stenosis divided by the quantity 100 minus true percent stenosis, varied with respect to the variables in question.

Performance will frequently be referred to as either 1) overestimation, in which case the mean estimate indicated more of a vessel constriction than the object of comparison (truth or human performance) or 2) underestimation, wherein the estimate indicated less constriction than the expected value (Figure 7.3).

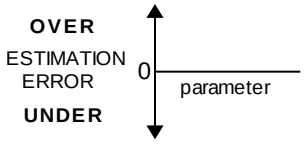


Figure 7.3. Stenosis over- and under- estimation. Positive errors reflect that the estimate indicates more vessel constriction than the expected value. Negative errors indicate underestimation relative to the standard.

The standard deviations reported are those about the mean for multiple observers. They are plotted as error bars in the graphs, and the bars are placed at plus and minus one standard deviation from the mean. Where the number of points in the plots is relatively small, tables of means and standard deviations are provided below the plots.

7.1.1 Analysis of Experiment Data

Human errors were studied as a function of the three factors in the experimental design: stenosis depth, EBS, and FNR. The levels and units of those variables are as described in Sections 4.3 and 6.1. As a reminder, the most noise occurs for low FNR's, while the most effective blur occurs for high EBS's. Data points at each condition are mean errors and standard deviations for 13 observers (Table 7.1).

DEPTH 25%			DEPTH 50%			DEPTH 75%		
FNR	EBS	mean	FNR	EBS	mean	FNR	EBS	mean
14.673	0.0031	0.018 (0.07)	7.966	0.0069	-0.059 (0.08)	2.258	0.1111	-0.037 (0.18)
12.636	0.0193	-0.012 (0.07)	6.504	0.0434	-0.055 (0.08)	1.806	0.1736	-0.088 (0.25)
13.047	0.0494	0.006 (0.06)	6.091	0.1111	-0.101 (0.14)	1.354	0.4444	-0.247 (0.33)
11.802	0.0031	-0.013 (0.08)	6.600	0.0069	-0.081 (0.12)	2.350	0.1111	-0.096 (0.18)
12.304	0.0193	0.010 (0.07)	6.984	0.0434	-0.144 (0.10)	1.948	0.1736	-0.257 (0.22)
13.024	0.0494	-0.002 (0.06)	5.605	0.1111	-0.132 (0.13)	0.780	0.4444	-0.295 (0.45)
10.326	0.0031	0.025 (0.07)	5.700	0.0069	-0.086 (0.08)	1.714	0.1111	-0.078 (0.21)
10.256	0.0193	-0.026 (0.06)	4.894	0.0434	-0.199 (0.11)	1.552	0.1736	-0.355 (0.31)
10.446	0.0494	0.010 (0.09)	5.057	0.1111	-0.061 (0.15)	0.958	0.4444	-0.149 (0.41)

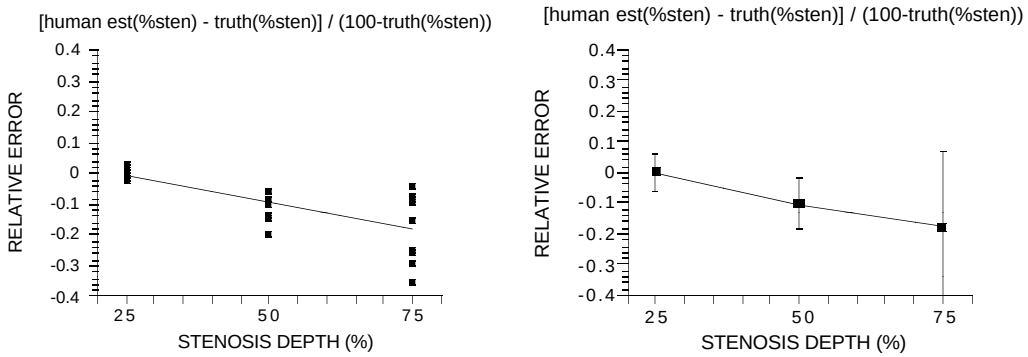
Table 7.1. Human stenosis estimation relative errors (and standard deviations) for all 27 depth, FNR, and EBS conditions. Means and standard deviations are for the 13 observers who participated at each condition.

The source table for the statistical analysis (Table 7.2) shows the probabilities for the main effects and interactions of the experimental factors. The statistics indicate that extent to which the independent variables listed caused significant variation in human relative error for stenosis estimation.

The three-way interaction of the angiography parameters was not statistically significant ($F=0.57$, $p=0.5651$), nor were any of the two-way interactions involving the stenosis depth factor. This suggests that the trends involving the other factors are similar at each stenosis depth and that it is therefore meaningful to examine independently the significant main effect due to stenosis depth ($F=34.40$, $p<0.0000$). Figure 7.4 demonstrates that observers underestimated, or indicated that the vessel was less constricted than it actually was, for the more constricted vessels. Accuracy was nearly perfect at the 25 percent depth condition.

SOURCE	F	Pr > F
depth	34.40	< 0.0001
EBS	9.47	0.0023
EBS•depth	0.22	0.8052
FNR	1.06	0.3048
FNR•depth	2.19	0.1132
EBS•FNR	4.87	0.0280
EBS•FNR•depth	0.57	0.5651

Table 7.2. Source table for the analysis of human stenosis estimation errors. The "SOURCE" column lists the main effects and interactions of the independent variables. Corresponding F statistics and probabilities indicate the statistical significance of the variation caused by the variables.



depth 25%	depth 50%	depth 75%
0.0018 (0.060)	-0.102 (0.084)	-0.178 (0.246)

Figure 7.4. Human stenosis estimation errors as a function of stenosis depth. At the left, the data at all 27 conditions is shown with a linear fit, while the right plots only the means with respect to stenosis depth.

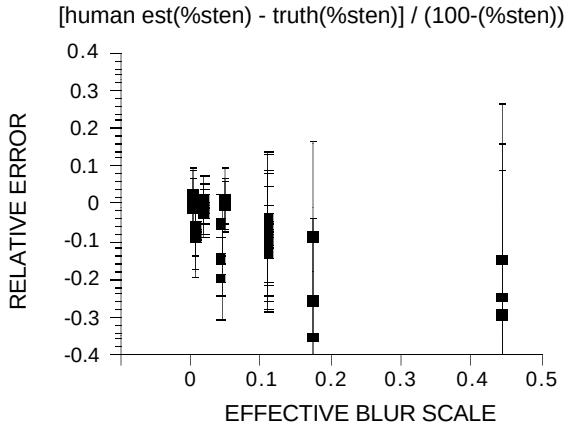


Figure 7.5. Human stenosis estimation errors as a function of effective blur scale.

The main effect due to the EBS parameter was also significant (F=9.47, p=0.0023). Although the EBS parameter to a large extent exhibits the trends that were captured by the stenosis depth main effect, the addition of the EBS term into the linear model reduced significantly further the error in the statistical model's representation of the data. All of the data are plotted as a function of the EBS factor in Figure 7.5. Generally the effect on human performance of an increase in the blur, as measured by the EBS parameter, is to cause larger underestimation errors.

There was no significant variation in human relative errors for the main effect of noise (F=1.06, p=0.3048). Although accuracy appears to vary with respect to FNR (Figure 7.6), the analysis indicates that the

residual error term in the linear model was not significantly reduced by the addition of an FNR factor. In Figure 7.6, the data are labeled according to their stenosis depths. The hypothesized differential effect of noise on stenoses of different depths may in part cause the variation in human accuracy as a function of depth (Figure 7.4). At the same time, since the FNR may reflect stenosis depth, the trends in accuracy as a function of FNR may have some relation to whatever the form of the depth trends.

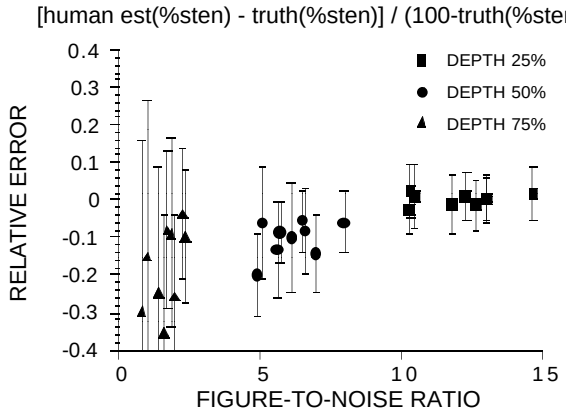
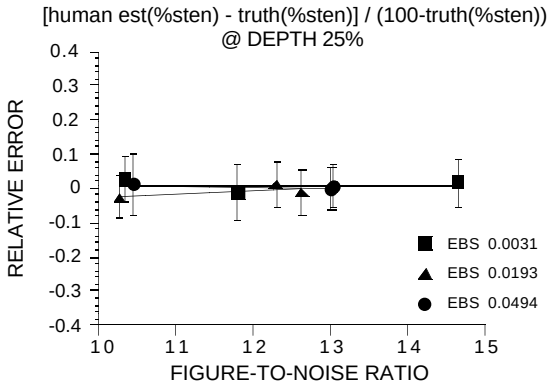


Figure 7.6. Human stenosis estimation errors as a function of figure-to-noise ratio.

The non-significant three-way interaction suggests that the blur-by-noise relationships have a similar form at each stenosis depth levels. However, there is no way to "combine" the data across the depth factor for presentation purposes: every condition is described by a unique FNR. The data are analyzed and plotted below separately at each of the stenosis depths (Figure 7.7 through 7.9). In each case the analysis source table with the EBS and FNR main effects and their interaction are provided beneath the plots. The plots also contain linear least squares fits through the data points belonging to a single EBS level. The conclusion represented by the non-significant three-way interaction is supported by the fact that the EBS-by-FNR interactions are the same (not significant) at every depth.

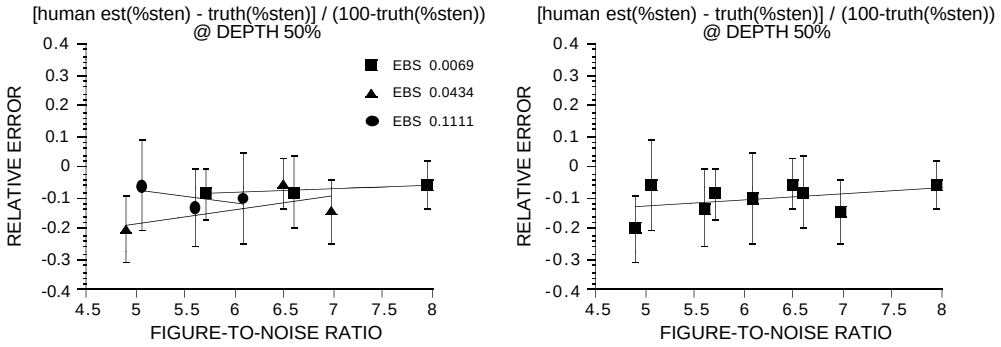


SOURCE	F	Pr > F
EBS	0.04	0.8426
FNR	0.53	0.4691
EBS*FNR	0.19	0.6660

Figure 7.7. Human stenosis estimation errors as a function of EBS and FNR at stenosis depth 25%.

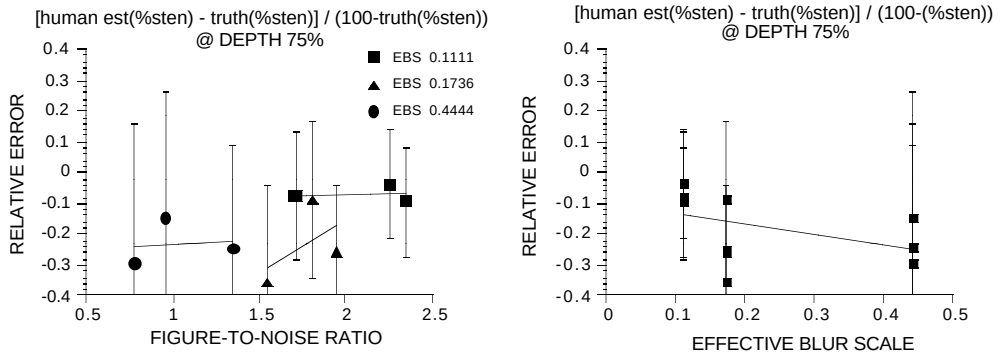
At the 25 percent stenosis depth condition, there were no significant main effects or an interaction: human accuracy is relatively constant and indeed good over the range of blur and noise conditions. There was, however, a marginally significant main effect of the FNR factor at the 50 percent stenosis depth condition ($F=5.28$, $p=0.0134$). In general underestimation errors decrease with increasing FNR (Figure 7.8, right). The data at that stenosis depth were not differentiated with respect to the EBS factor ($F=0.43$, $p=0.5117$). Finally, at the

most constricted stenosis condition, stenosis depth 75 percent, it is the blurring that causes the variation in estimation accuracy (F=7.04, p=0.0091).



SOURCE	F	Pr > F
EBS	0.43	0.5117
FNR	5.28	0.0134
EBS*FNR	3.42	0.0671

Figure 7.8. Human stenosis estimation errors as a function of EBS and FNR at stenosis depth 50%. Linear fits for the three EBS conditions are shown in the plot at the left. The same data are shown on the right with a linear fit through all nine conditions.



SOURCE	F	Pr > F
EBS	12.71	0.0091
FNR	2.84	0.0945
EBS*FNR	3.64	0.0590

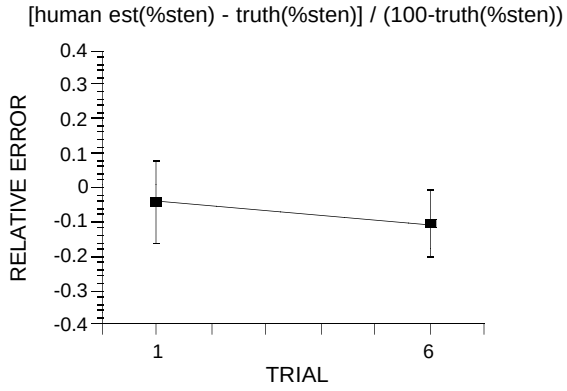
Figure 7.9. Human stenosis estimation errors as a function of EBS and FNR at stenosis depth 75%. Linear fits for the three EBS conditions are shown as a function of FNR in the plot at the left. The same data are shown on the right as a function of EBS with a linear fit through all nine conditions.

Human performance in the angiography experiment may be summarized as follows. The effects of blur and noise were as expected: inspection of the data shows relative errors in general increased with more blur and noise (Figures 7.5 and 7.6). Also, accuracy worsened with increased percent stenosis (Figure 7.4). Humans tended to underestimate vessel depths at the 50 and 75 percent stenosis depth conditions. It was only for mildly constricted vessels that overall accuracy was excellent and relatively unaffected by the noise and blur conditions. These trends in human performance with respect to stenosis depth are discussed further in Section 7.1.3, where accuracy was measured in the absence of blur and noise.

The statistical significance of many of the effects furthermore suggests general success in achieving human variation along the parameter levels chosen for the study. That the statistical effects that say something about the suitability of the parameter level choices were not entirely significant may be further evidence for the superiority of a description of the imaging parameters via the perceptual measures EBS and FNR. A set of experimental conditions that were described by equal increments in those measures might have produced more consistent variation in human estimation accuracy. The discussion (Section 7.6.4) reflects further upon the appropriateness of these measures.

7.1.2 Practice Analysis

The human data were analyzed to test for the possibility of practice or fatigue effects. The observers were fairly naive to the particular task in this study and, in spite of the practice at the beginning, might have improved over the course of the experiment. This analysis examined whether each observer's error at the chronologically first trial in each condition was significantly different from the same measure at the last trial.



1	6
-0.042 (0.119)	-0.105 (0.096)

Figure 7.10. Human stenosis estimation errors at the first (1) and last (6) trial across all angiographic conditions.

There was a practice (or fatigue) effect on human performance. There was no significant four-way interaction ($F=0.62$, $p=0.5381$); whatever the three-way trends among stenosis depth, EBS, and FNR were, they were similar at the first and last trials. There were also no other statistically significant higher order interactions involving the trail variable. However, there was a significant main effect of the trial variable alone ($F=6.94$, $p=0.0086$). Figure 7.10 demonstrates that the overall accuracy at the last trial in every condition was worse than at the first. Apparently the practice provided prior to the experiment, together with the random ordering of images per observer, was not entirely successful in preventing undesirable drifts in overall accuracy during the study. The effects may simply be fatigue. More likely the result is a drift in overall observer bias due to practice and exposure that amounted to further underestimation.

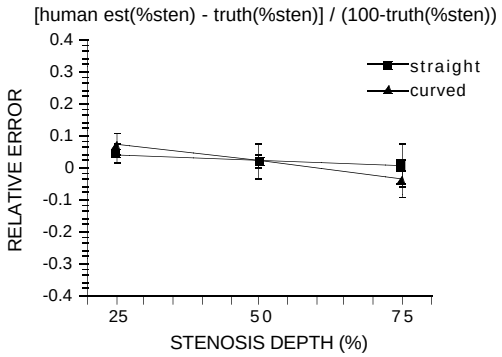
As it stands, the practice analysis that was performed here is particularly suspect due to the "fidget factor:" observers often perform poorly towards the end of the experiment as they anticipate the termination of their duties. However, the design of the angiography experiment made it difficult to conclusively test for the presence of practice or fatigue effects by any other means other than that just described. Had this problem been anticipated, the trials in the experiment could have been counterbalanced by "block" such that a trial from each of the 27 experimental conditions was conducted before moving on to another such block. This design would have allowed an independent variable, "block," as part of the analysis, and significant trends in that variable would be indicative of practice effects over the course of the experiment.

7.1.3 Plain Vessels

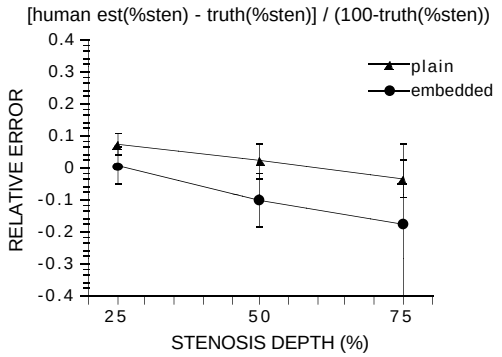
Some indication of baseline performance for stenosis estimation was desired as additional information in understanding and comparing human and model performance. Human performance for "plain" vessels placed on a black background with no noise or blur was examined in a separate experiment with six graduate student observers. Straight vessels like the measuring tool prototypes, as well as the 27 multiply curved vessels used in the experiment images, were generated with each of the three stenosis depths (25, 50, and 75 percent). Ten observations were collected from each of the observers for each of the three straight vessel depths, while one observation was collected for each of the ($27*3$) curved vessels. All other aspects of the experimental protocol, including the width variation of the prototype vessel from trial to trial, remained identical to the main study. An ANOVA with depth as the single fixed factor was used to analyze the results.

Human accuracy for estimation of the straight vessels was unaffected by the different depths ($F=0.82$, G-G $p=0.4147$). However, performance did change as a function of depth for curved vessels ($F=12.45$, G-G $p=0.0036$). For these plain vessels with wandering paths, observers overestimated the stenosis depth with a relative error of approximately 7% at the mildest constriction (Figure 7.11). This is in contrast to human performance with the same vessels embedded in backgrounds, noise, and blur, where the previously discussed

main effect suggested most accurate performance at the smallest stenosis depth and increasing underestimation in the degraded conditions. Thus performance appears to be simply shifted: the imposition of the experimental noise and blur conditions caused about a 0.10 shift in relative error toward stenosis underestimation.



	depth 25%	depth 50%	depth 75%
straight	0.039 (0.030)	0.020 (0.022)	0.005 (0.068)
curved	0.069 (0.035)	0.018 (0.052)	-0.040 (0.059)



	depth 25%	depth 50%	depth 75%
embed	0.002 (0.060)	-0.102 (0.084)	-0.178 (0.246)
plain	0.069 (0.035)	0.018 (0.052)	-0.040 (0.059)

Figure 7.11. Human stenosis estimation errors for plain vessels. At the left, errors are plotted for straight and curved plain vessels. At the right, human performance for the plain curved vessels is superimposed on the experiment ("embedded") vessel data from Figure 7.4.

Human accuracy as a function of stenosis depth for the plain straight vessels exhibits something like the zoom invariance predictions. That is, across the stenosis depth conditions relative errors are roughly constant. It is unclear why errors were not constant for plain curved vessels. For the vessels that were embedded in the degraded backgrounds, some amount of shift results, and more importantly, accuracy is worse and uncertainty is highest at the smallest width conditions. This result may be due to the detrimental effect on accuracy and certainty that noise and blur would be expected to have at the smaller width conditions.

7.2 Model Performance for Stenosis Estimation

The model's performance for stenosis estimation is examined in this section. Similar to the previous discussion, the outcome of interest is the model's relative error, or the model's stenosis estimate minus the true stenosis percent divided by 100 minus the true stenosis percent. Plots of model error are shown as a function of the same experimental variables. The value plotted at each condition is the mean percent stenosis estimation error over the six backgrounds that served in the condition.

No statistical conclusions are offered in this section. First, there is only a single mean (across the six backgrounds) from the model at each condition. A statistical analysis uses the mean and variance from *multiple* observers at each condition to make a conclusion about whether the means vary across the conditions. Second, any statistical analysis tests the hypothesis that the difference between two or more observations occurred due to chance alone. However, for this same set of experimental images, the model would produce the same results again.

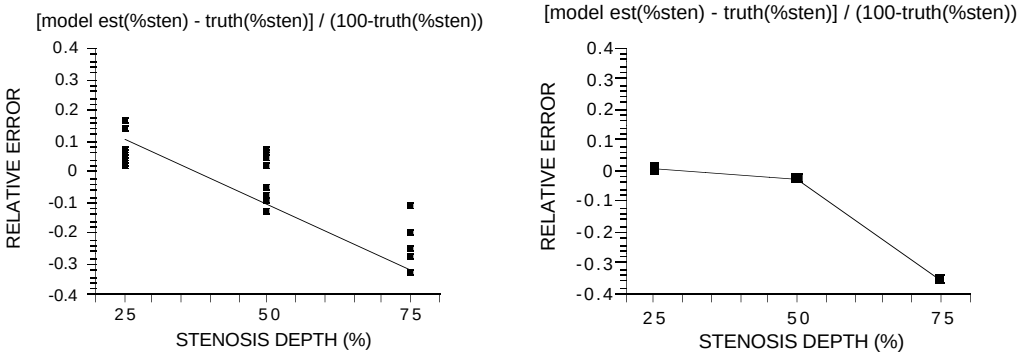
Similarly, until the theoretical investigations reported in Section 7.3.3, no standard deviations are reported for the model. An "intra-" or "inter-observer" variability could not be reported for the model for the experimental data, since there is only one model and it produces the same estimate every time it operates upon the same image. Nonetheless, the visible trends in model errors are informative.

7.2.1 Analysis of Experiment Data

The entire data set for model stenosis estimation errors is in Table 7.3. Figure 7.12 shows the model results for the main effect of stenosis depth. This parameter clearly has a big influence on model errors: the model overestimates the depth of mildly constricted vessels but underestimates more severely constricted vessels.

DEPTH 25%			DEPTH 50%			DEPTH 75%		
FNR	EBS	mean	FNR	EBS	mean	FNR	EBS	mean
14.673	0.0031	0.055	7.966	0.0069	-0.051	2.258	0.1111	-0.114
12.636	0.0193	0.064	6.504	0.0434	0.075	1.806	0.1736	-0.520
13.047	0.0494	0.068	6.091	0.1111	-0.123	1.354	0.4444	-0.460
11.802	0.0031	0.036	6.600	0.0069	0.062	2.350	0.1111	-0.275
12.304	0.0193	0.017	6.984	0.0434	-0.076	1.948	0.1736	-0.327
13.024	0.0494	0.030	5.605	0.1111	-0.089	0.780	0.4444	-0.250
10.326	0.0031	0.142	5.700	0.0069	0.044	1.714	0.1111	-0.197
10.256	0.0193	0.017	4.894	0.0434	-0.092	1.552	0.1736	-0.541
10.446	0.0494	0.170	5.057	0.1111	0.019	0.958	0.4444	-0.502

Table 7.3. Model stenosis estimation relative errors for all 27 depth, FNR, and EBS conditions. Means are with respect to the six angiographic backgrounds that served in each condition.



depth 25%	depth 50%	depth 75%
0.067	-0.026	0.354

Figure 7.12. Model stenosis estimation errors as a function of stenosis depth. At the left, the data from all conditions are plotted with a linear fit. On the right, the means at each depth are indicated.

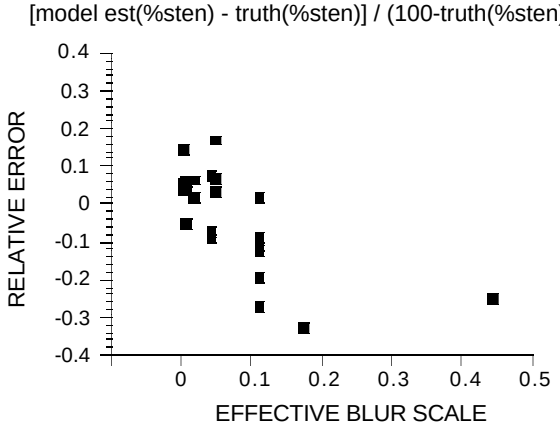


Figure 7.13. Model stenosis estimation errors as a function of effective blur scale.

The main effect for blur also shows a great variation in model accuracy as a function of the EBS factor (Figure 7.13).

The last main effect is that for noise, where overall the model is least accurate at the lowest FNR conditions (Figure 7.14). Again the trends across EBS and FNR capture the variation in accuracy demonstrated as a function of the depth parameter.

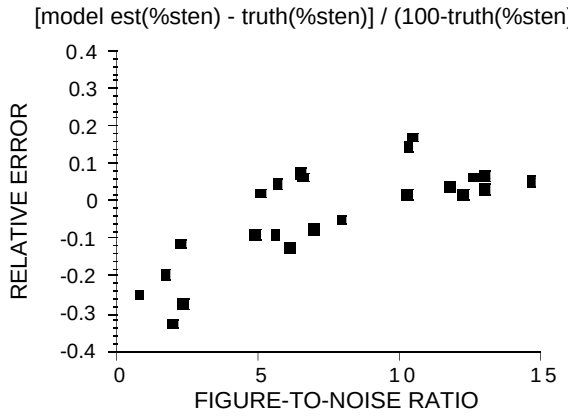


Figure 7.14. Model stenosis estimation errors as a function of figure-to-noise ratio.

The following graphs (Figures 7.15 through 7.17) plot the results as a function of the noise and blur parameters for the stenosis depths separately.

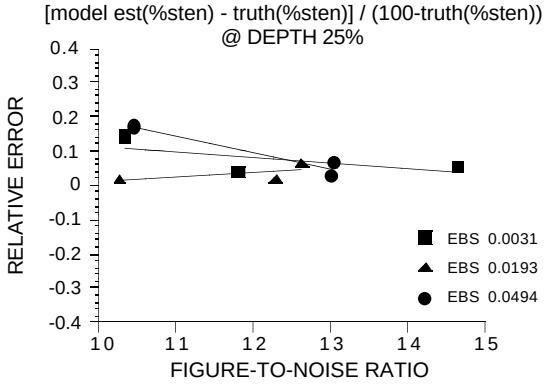


Figure 7.15. Model stenosis estimation errors as a function of EBS and FNR at stenosis depth 25%.

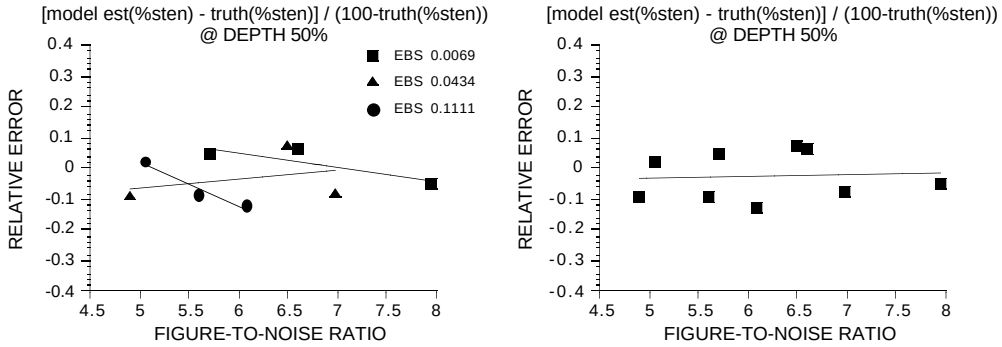


Figure 7.16. Model stenosis estimation errors as a function of EBS and FNR at stenosis depth 50%. At the right, a linear fit through all points is shown.

The variations in the experimental design parameters, stenosis depth and the noise and blur descriptors, did as intended result in variation in model performance. For instance, the model greatly overestimates the severity of the vessel constrictions in the degraded conditions for the vessels with mild (25 percent) stenoses but performs quite accurately at some of the milder blur and noise conditions (Figure 7.15). And Figure 7.18 (right) shows the trend of overall improvement that results from a decrease in the EBS. However, sometimes the variation in the model's behavior is not always understandable or ordered with respect to the parameters. There

are for example two instances at the two larger stenosis depths (50 and 75 percent) where accuracy for a fixed EBS diminishes with an increase in the FNR.

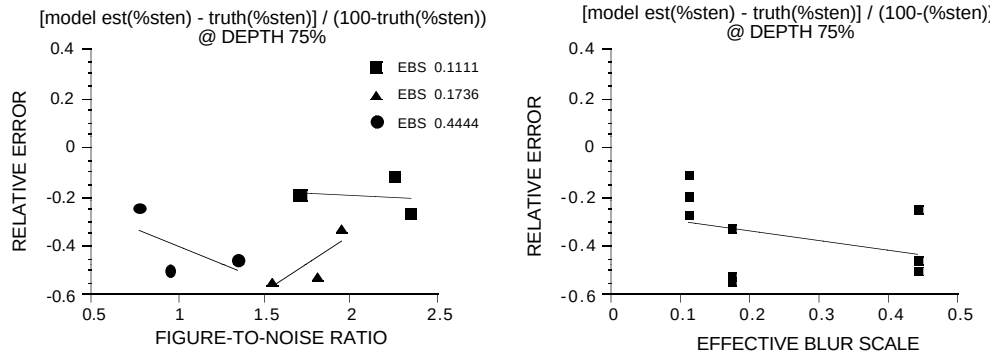
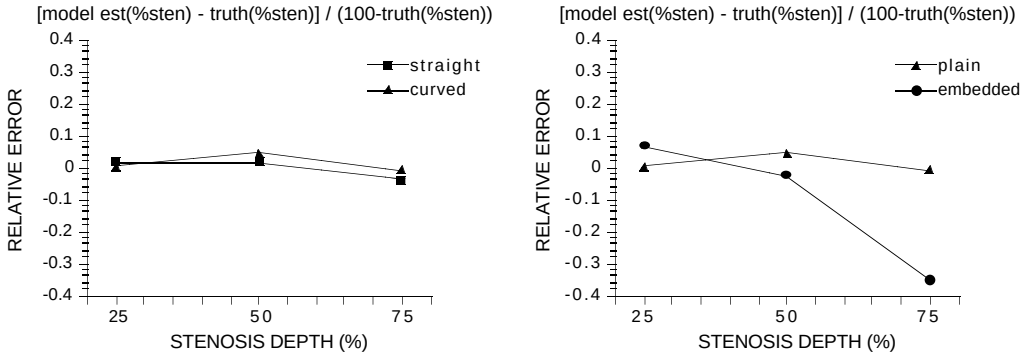


Figure 7.17. Model stenosis estimation errors as a function of EBS and FNR at stenosis depth 75%. At the right, a linear fit through all nine points is shown as a function of EBS.

7.2.2 Plain Vessels

Model accuracy for the same straight and curved plain vessels as those discussed in Section 7.1.3 above is relatively constant and quite good (Figure 7.18, left). This constant performance is very different from the model's accuracy for these stenosis depths in the presence of backgrounds, blur, and noise. In those experimental conditions, model performance ranged from overestimation errors of 0.067 at the 25 percent stenosis to underestimation errors of 0.354 at 75 percent depth condition (Figure 7.18, right).



	depth 25%	depth 50%	depth 75%
straight	0.015	0.015	-0.040
curved	0.0024	0.043	-0.0095

	depth 25%	depth 50%	depth 75%
embed	0.067	-0.026	0.354
plain	0.0024	0.043	-0.0095

Figure 7.18. Model stenosis estimation errors for plain vessels. Model errors are plotted for straight and curved plain vessels (left). At the right, the plain curved vessel performance is replotted together with the stenosis depth main effect results from Figure 7.14.

7.3 Human vs. Model Performance for Stenosis Estimation

This section contains a number of comparisons of model and human performance for the stenosis estimation task. The question is whether the model behavior presented in the previous section parallels the human results described in Section 7.1. To the extent that the results are not perfectly comparable in all cases, it will be of interest to ascertain the conditions and parameters under which the model was and was not sufficiently predictive. The outcome measure for comparing human and model stenosis estimation accuracy is the difference between model and human relative percent stenosis errors at each of the 27 experimental conditions. All plots and analyses in this section utilize this measure.

7.3.1 Analysis of Experiment Data

The full results are shown in Table 7.4. The source table for the statistical analysis is Table 7.5. The highest order interaction was not statistically significant (F=1.82, p=0.1640). The lower order interactions can thus be considered. There was no interaction of the depth parameter with FNR (F=0.66, p=0.5165), nor was there an interaction of depth with EBS (F=1.36, p=0.2580). Thus, it is reasonable to examine the main effect of stenosis depth alone (Figure 7.19).

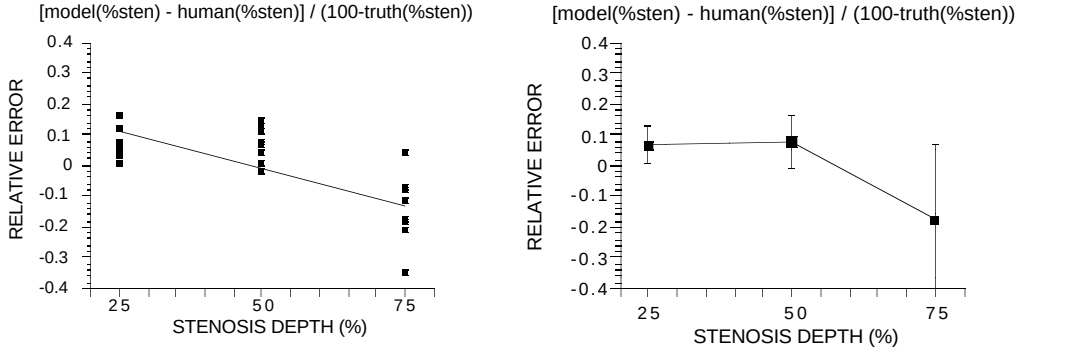
DEPTH 25%			DEPTH 50%			DEPTH 75%		
FNR	EBS	mean (sd)	FNR	EBS	mean (sd)	FNR	EBS	mean (sd)
14.673	0.0031	0.037 (0.07)	7.966	0.0069	0.008 (0.08)	2.258	0.1111	-0.077 (0.18)
12.636	0.0193	0.077 (0.07)	6.504	0.0434	0.130 (0.08)	1.806	0.1736	-0.431 (0.25)
13.047	0.0494	0.062 (0.06)	6.091	0.1111	-0.023 (0.14)	1.354	0.4444	-0.213 (0.33)
11.802	0.0031	0.049 (0.08)	6.600	0.0069	0.143 (0.12)	2.350	0.1111	-0.179 (0.18)
12.304	0.0193	0.007 (0.07)	6.984	0.0434	0.068 (0.10)	1.948	0.1736	-0.069 (0.22)
13.024	0.0494	0.031 (0.06)	5.605	0.1111	0.043 (0.13)	0.780	0.4444	-0.045 (0.45)
10.326	0.0031	0.117 (0.07)	5.700	0.0069	0.131 (0.08)	1.714	0.1111	-0.118 (0.21)
10.256	0.0193	0.043 (0.06)	4.894	0.0434	0.107 (0.11)	1.552	0.1736	-0.186 (0.31)
10.446	0.0494	0.160 (0.09)	5.057	0.1111	0.080 (0.18)	0.958	0.4444	-0.352 (0.41)

Table 7.4. Model-human relative error differences for stenosis estimation. Means and standard deviations for the 13 difference scores at the 27 depth, FNR, and EBS conditions are shown.

SOURCE	F	Pr > F
depth	76.08	< 0.0001
EBS	0.55	0.4597
EBS•depth	1.36	0.2580
FNR	5.89	0.0157
FNR•depth	0.66	0.5165
EBS•FNR	8.87	0.0031
EBS•FNR•depth	1.82	0.1640

Table 7.5. Source table for the analysis of model-human stenosis estimation differences.

Both the human and the model results were shown previously (Figures 7.4 and 7.12) to exhibit the same trend as a function of stenosis depth: overestimation occurred for the mild stenoses while underestimation resulted for severe constrictions. However, because the model does this to a greater extent, the difference in performance is not constant.



depth 25%	depth 50%	depth 75%
0.065 (0.060)	0.076 (0.084)	-0.176 (0.246)

Figure 7.19. Model-human stenosis estimation differences as a function of stenosis depth.

Table 7.5 indicates a significant interaction of the EBS and FNR parameters (F=8.87, p=0.0031). The non-significant three-way interaction suggests that the form of this EBS-by-FNR interaction is roughly the same at each stenosis depth condition. Thus to illustrate the two-way interaction, model-human relative error differences

were averaged across the three depth conditions. This also required an averaging of the independent parameters, EBS and FNR, describing the three conditions that were averaged in each case. The linear fits in Figure 7.20 suggest that the interaction is caused by the conditions at the highest level of blur, or EBS 0.2018, where the linear trend indicates, contrary to the other two blur conditions, increasingly less underestimation by the model relative to humans as a function of FNR.

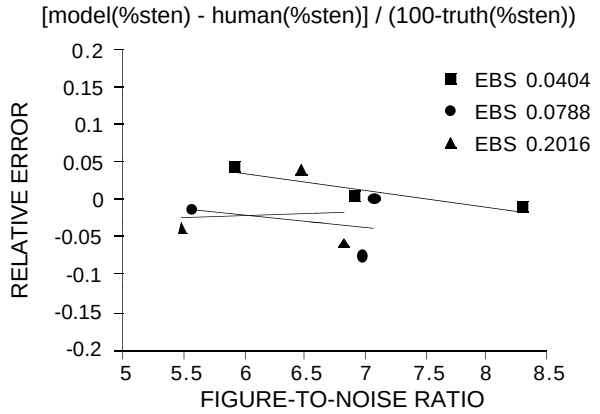


Figure 7.20. Model-human stenosis estimation differences as a function of FNR. Linear fits through the three EBS conditions are shown. The nine data points, and the FNR and EBS levels describing them, are means across the three stenosis depth conditions.

Although there did exist the EBS-by-FNR interaction described previously, it is useful to look at the data as functions of EBS and FNR alone. The trend as a function of increasing EBS (Figure 7.21) is for model underestimation relative to humans. Figure 7.22 demonstrates relatively constant, small differences between the model and humans at the intermediate and high FNR conditions. At the lowest FNR conditions, the differences are largest and most variable.

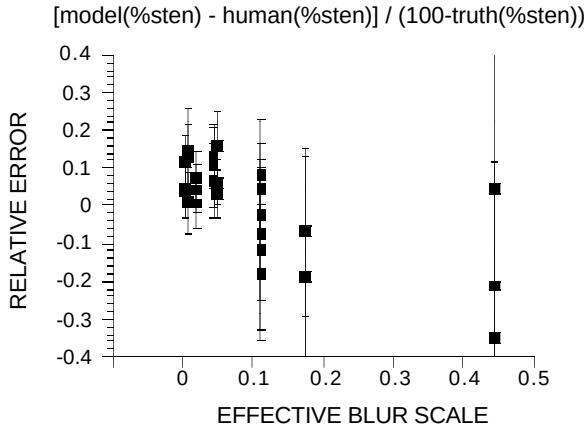


Figure 7.21. Model-human stenosis estimation differences as a function of effective blur scale.

Figures 7.23 through 7.26 plot the blur and noise data at each stenosis depth. The non-significant three-way interaction suggests that the significant blur-by-noise interaction ($F=8.87$, $p=0.0031$) takes on essentially the same form at each depth. Source tables are provided with the plots in each case. Again, linear least-squares fits are shown.

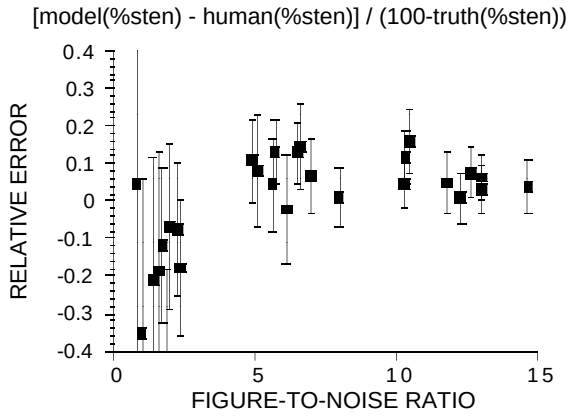
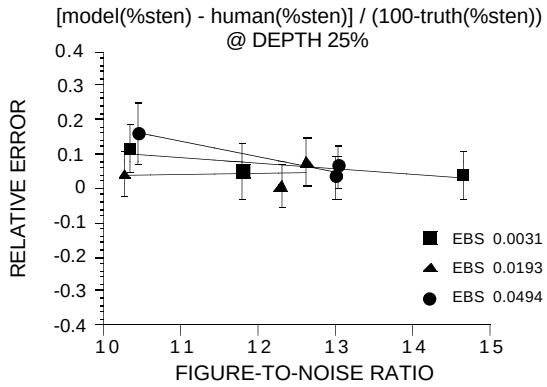


Figure 7.22. Model-human stenosis estimation differences as a function of figure-to-noise ratio.

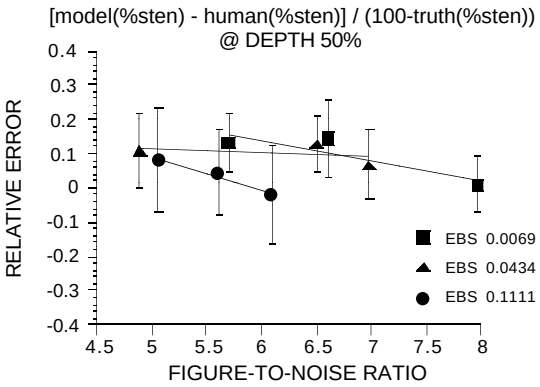
The large overestimation errors for the model in the more degraded blur and noise conditions at the 25 percent stenosis depth contrasted strongly with the accurate human performance there. In the less degraded conditions at the 25% depth, the differences in the relative errors were more constant and for the most part roughly less than 0.10. These trends result in a significant interaction (F=10.76, p=0.0014).



SOURCE	F	Pr > F
EBS	4.05	0.0465
FNR	27.61	< 0.0001
EBS*FNR	10.76	0.0014

Figure 7.23. Model-human stenosis estimation differences as a function of FNR and EBS at stenosis depth 25%. Linear fits for the three EBS conditions are shown as a function of FNR.

The non-significant interaction at the 50 percent depth condition (Figure 7.24, F=0.06, p=0.8126) permits examination of the main effects (Figure 7.25). For both the EBS and FNR parameters, the plots show aptly that the differences were not constant. The variation in the differences occurs because the model's accuracy is generally constant as a function of FNR at the 50 percent depth condition (Figure 7.16, right) whereas human accuracy is more variable (7.8, right). In the EBS plot (Figure 7.25, right), the difference between model and human errors was actually nearest zero in the highest blur conditions.



SOURCE	F	Pr > F
EBS	12.05	0.0007
FNR	18.42	< 0.0001
EBS*FNR	0.06	0.8126

Figure 7.24. Model-human stenosis estimation differences as a function of FNR and EBS at stenosis depth 50%.

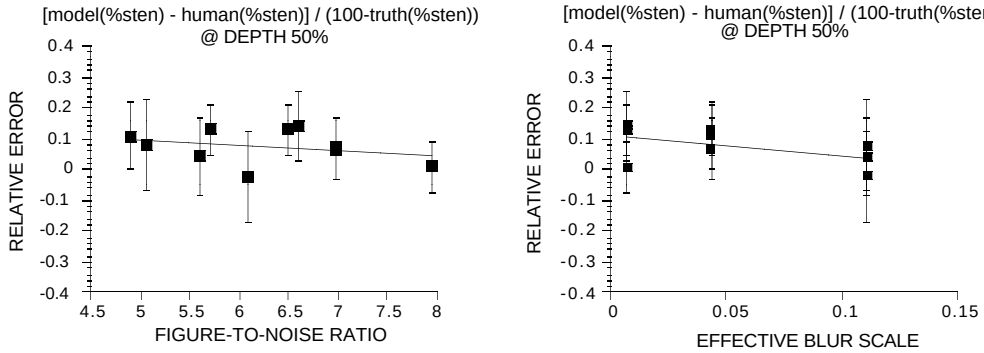
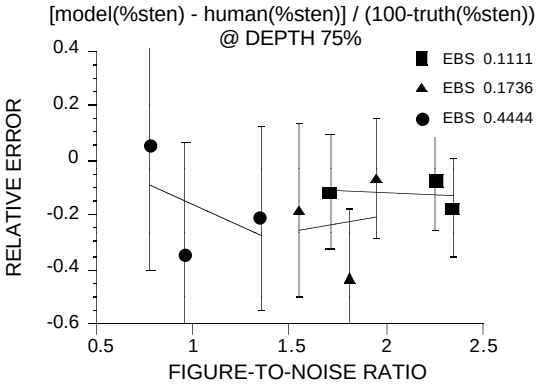


Figure 7.25. Model-human stenosis estimation differences as a function of FNR (left) and EBS (right) at stenosis depth 50%.



SOURCE	F	Pr > F
EBS	0.08	0.7768
FNR	0.39	0.5319
EBS*FNR	7.37	0.0077

Figure 7.26. Model-human stenosis estimation differences as a function of EBS and FNR at stenosis depth 75%.

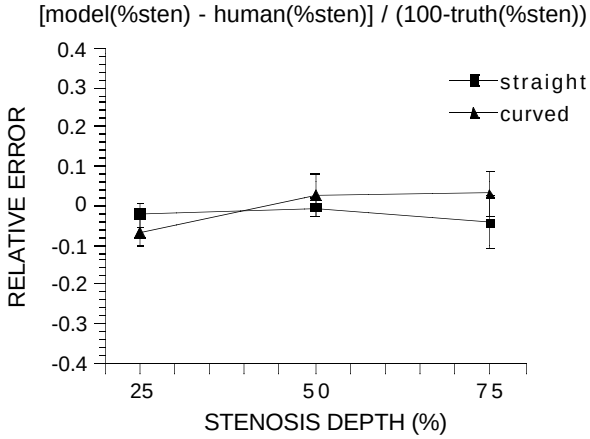
Finally, at the 75 percent stenosis condition (Figure 7.26), there is a significant interaction (F=7.37, p=0.0077). The main effects for human and model accuracy alone (Figures 7.9 and 7.17) both show decreasing accuracy with an increase in the EBS parameter. Indeed, the statistics indicate no significant main effects for the relative errors for both the EBS and FNR parameters. Variation in the errors appears to result from simultaneous changes in both blur and noise.

7.3.2 Plain Vessels

Relative errors for the comparison of model and human estimates for the plain vessels are shown in Figure 7.27. The differences for the straight vessels are constant and nearly zero. The statistical analysis indicated no significant variation in the scores across the three depths (F=1.13, G-G p=0.3393). In that case, both the model and humans performed the task with consistent and good accuracy to begin with (Figures 7.11 and 7.18).

For the curved vessels, the larger difference between the model and the human data at the 25 percent stenosis condition is not consistent with the differences at the other two depths, and the overall effect of stenosis depth was statistically significant (F=12.50, G-G p=0.0035). From Section 7.13 it is known that humans overestimated at the 25 percent condition, while their accuracy at the other two depths was nearly perfect. The model's flat performance for all three depths thus generates the disparity at the 25 percent stenosis depth condition.

It could be hypothesized that the discrepancies between the model and human results as a function of stenosis depth were due in part to the failure of the model used in these studies to incorporate the lack of perfect zoom invariance that humans have been shown to exhibit. The model results do seem to exhibit the zoom invariance characteristics. In the absence of blur and noise, the model's relative errors are roughly constant, while the human errors are not.



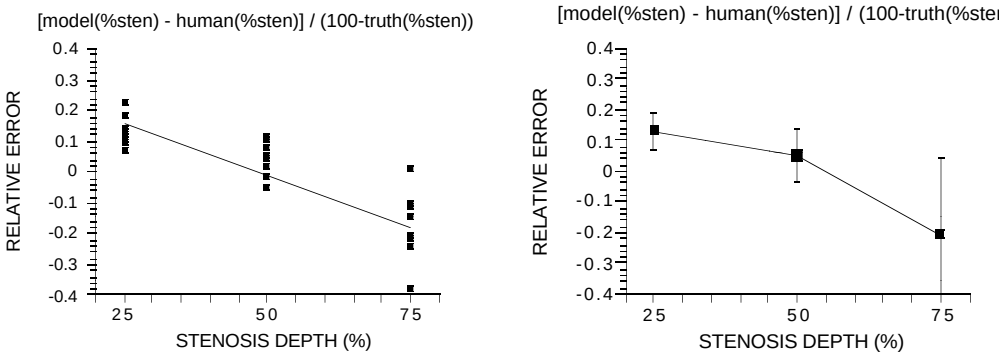
	depth 25%	depth 50%	depth 75%
straight	-0.023 (0.030)	-0.0049 (0.021)	-0.045 (0.068)
wiggle	-0.067 (0.035)	0.025 (0.052)	0.030 (0.059)

Figure 7.27. Model-human stenosis estimation differences for plain vessels.

When viewing the experiment data, where blur and noise were present, one must consider in the interpretation that blur and noise may work against the zoom invariance principles. Because we believe that perceptual effects of blur and noise are dependent on object size, the degradations used in the experiment might be expected to have, on the whole, an increasingly detrimental effect on accuracy as stenotic width is decreased. This is opposite from the constant trend in relative error predicted by zoom invariance. This is indeed what appears to have occurred with the human data. Figure 7.4 demonstrates the increase in errors that occur for more constricted vessels. Yet peak accuracy for the model for the experiment images occurs roughly at the 50 percent stenosis depth condition.

Thus, while to some extent the disparity between the model and the human in the experiment data may be due to the different properties with respect to zoom, it is clear that the experimental conditions themselves contributed significantly to the model and human differences. The way that the model performance changes relative to plain vessels when it was subjected to blur and noise is very different from the simple shift in the human results when moving from plain to embedded vessels (Figures 7.11 and 7.18). This observation is one more indication that model does not handle or respond to blur and noise in the same way that humans do.

To examine the impact of the baseline differences between model and human accuracy, the model stenosis estimates for the experiment images were adjusted by the amount by which they were different from the human estimates for the plain vessels. That is, the "curved" plot from Figure 7.27 was subtracted from the plot in Figure 7.19. The result (Figure 7.28) is that the stenosis estimation differences as a function of stenosis depth are nearly linear and vary quite substantially.



depth 25%	depth 50%	depth 75%
0.132 (0.060)	0.051 (0.084)	-0.206 (0.246)

Figure 7.28. Model-human stenosis estimation differences as a function of stenosis depth. Model-human differences at each depth were adjusted by the model-human differences at the plain wiggle vessels. The plot on the left shows a linear fit through all 27 conditions, while the right shows the mean and standard deviation of the differences at each depth.

While this adjustment seems to clarify the differences between the model and the human as a function of stenosis depth, it does not improve the correspondence between the two. This was true when viewing the results as a function of the EBS and FNR parameters, and nothing further of this adjustment is shown.

7.3.3 Comparison of Standard Deviations

Another model-based estimate that might be used as a metric for image quality is the variance in the accuracy of a task. Not only is the mean accuracy with which a clinical estimate is determined important, but the certainty and consistency of the estimate may be indicative of goodness.

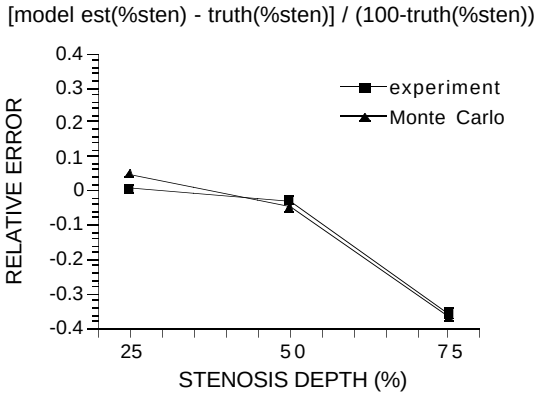
There were, in addition to the independent variables studied, three sources of variability in the human estimates in the angiography study. First, there were six backgrounds randomly assigned to each of the experimental conditions. There is thus a standard deviation at each condition for each observer that is computed about the mean for the observations over the six backgrounds in the conditions. Second, the scores at each condition for the overall analysis were means across thirteen observers, and there is an inter-observer standard deviation about each of those means. Finally, human observers possess an internal, or intra-observer, variability owing to the inherent uncertainty exhibited by all measurements made by our senses.

The only variability that can be estimated for the *model* from the *experiment* data is that due to the background variation. A first analysis was conducted to test whether the difference between the human and model standard deviations in relative error due to background vary across the conditions. Table 7.6 is the source table for the repeated covariates regression analysis. There was a significant three-way interaction ($F=8.72$, $p=0.0002$), as well as other highly significant interactions. Clearly the model's standard deviations in stenosis estimation with respect to the backgrounds in each condition did not parallel those from the humans.

SOURCE	F	Pr > F
depth	39.00	< 0.0000
EBS	34.31	< 0.0000
EBS•depth	0.53	0.5904
FNR	0.37	0.5423
FNR•depth	14.15	< 0.0000
EBS•FNR	53.59	< 0.0000
EBS•FNR•depth	8.72	0.0002

Table 7.6. Source table for the analysis of model-human stenosis estimation standard deviation differences due to background variation.

What is really needed, though, is an approximation, assigned to the model, that can be its intra- and inter-observer variability. This estimate for the "natural" or inherent variability for the model was determined by measuring its performance for many iterations, or realizations, of the noise at each condition. Specifically, for each of the backgrounds at each of the 27 conditions, stenosis estimates were computed according to the core-based protocol (Section 4.4) for 75 iterations of the noise level assigned to that condition. The standard deviation about the mean relative error for the 6*75 stenosis estimates was designated as the model's deviation at each condition. Several results of this Monte Carlo simulation are described next.



	depth 25%	depth 50%	depth 75%
exp	0.0067	-0.026	-0.354
MC	0.051	-0.047	-0.365

Figure 7.29. Experimental and Monte Carlo model stenosis estimation errors as a function of stenosis depth.

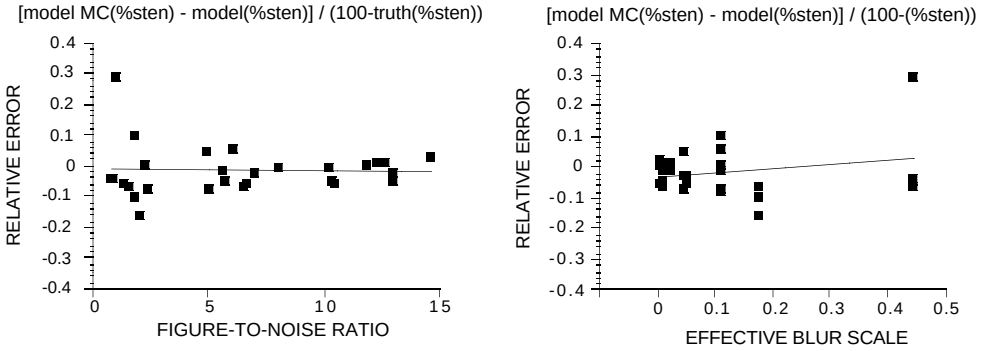


Figure 7.30. Stenosis estimation differences between Monte Carlo and experimental model results.

Before proceeding to an analysis of the standard deviations, the mean stenosis estimation errors from the Monte Carlo simulation were compared with the model errors for the images used in the experiment. Figure 7.29 compares the Monte Carlo and experiment main effects for stenosis estimation errors as a function of stenosis depth, and clearly the agreement is excellent. Differences and linear fits to those differences are plotted in Figure 7.30 as a function of FNR and EBS. Although there are differences between the estimates, the overall agreement is good.

The comparisons of the Monte Carlo means with the model's results for the experiment images alone provide some assurance that the performance of the model for the single realization of the noise in each image in the experiment was reasonably representative of its overall behavior for the noise conditions. However, there were enough differences between the stenosis estimation errors for the Monte Carlo and experiment data at the individual conditions to warrant a test of the Monte Carlo estimation errors against the human errors. The mean errors from the simulation, because they reflect so many more observations, may represent more appropriately the model's performance for the conditions in the experiment. These mean errors were tested against the human mean errors with a repeated covariates regression analysis in the same way that the original model results were.

The source table (Table 7.7) suggests substantial variation in the difference scores. The significant three-way interaction (F=7.66, p=0.0006) leads to the generation of Table 7.8, which shows the blur and noise trends at each stenosis extent. There are at the 50 and 75 percent conditions several significant main effects and interactions. Thus the Monte Carlo stenosis estimation errors were not any more successful at predicting the human errors than were the model errors from the original experiment.

SOURCE	F	Pr > F
depth	74.76	< 0.0000
EBS	3.60	0.0588
EBS•depth	1.68	0.1877
FNR	3.89	0.0494
FNR•depth	4.71	0.0096
EBS•FNR	19.98	< 0.0000
EBS•FNR•depth	7.66	0.0006

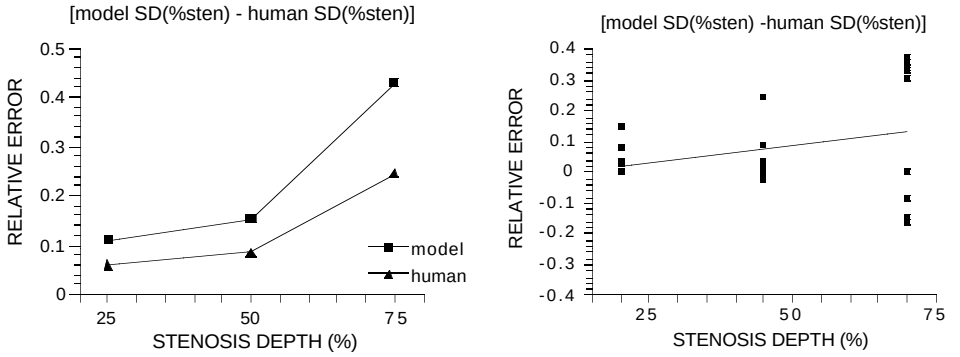
Table 7.7. Source table for the analysis of Monte Carlo model and human stenosis estimation differences.

	DEPTH 25%		DEPTH 50%		DEPTH 75%	
SOURCE	F	Pr > F	F	Pr > F	F	Pr > F
EBS	4.42	0.0377	4.87	0.0294	3.63	0.0594
FNR	6.09	0.0151	19.98	< 0.0001	5.55	0.0202
EBS*FNR	26.71	< 0.0001	5.02	0.0271	20.97	< 0.0001

Table 7.8. Source table for the analysis of Monte Carlo model and human stenosis estimation differences at the three stenosis depth conditions.

Next, the model standard deviations from the Monte Carlo simulation were compared with the standard deviations about the mean relative error for the thirteen observers. No statistical analyses were attempted on these data: the non-orthogonality of the experimental design and the lack of intra-observer variability estimates would make any analysis in this situation difficult. Nonetheless the plots are useful in understanding the trends.

Figure 7.31 shows the main effect for the standard deviation differences for stenosis depth. At the left, the parallelism of the standard deviations combined across the blur and noise conditions is good. However, the plot at the right in Figure 7.31 suggests that the differences are most variable in the highly-constricted vessel conditions.



	depth 25%	depth 50%	depth 75%
model	0.110	0.153	0.430
human	0.060	0.084	0.246

Figure 7.31. Human and model standard deviations as a function of stenosis depth. At the left, Monte Carlo estimates for the model and human standard deviations combined across other experimental conditions are plotted versus depth. A linear fit through the standard deviation differences is shown at the right.

The linear fits through the deviation differences plotted against FNR (Figure 7.32) and EBS (Figure 7.33) demonstrate some constancy as well. However, there are substantial differences between the model and human standard deviations in the highest blur and noise conditions.

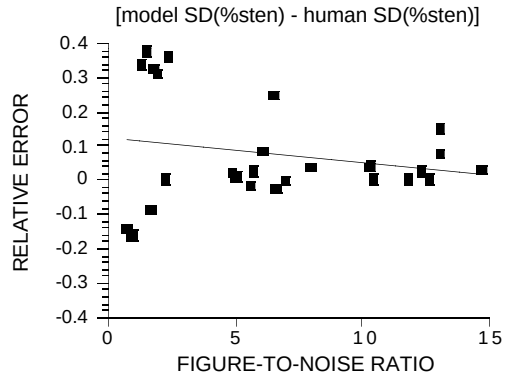


Figure 7.32. Model-human standard deviation differences as a function of figure-to-noise ratio.

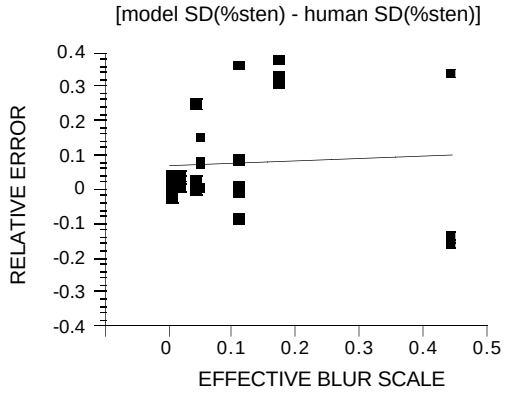


Figure 7.33. Model-human standard deviation differences as a function of effective blur scale.

Figures 7.34 through 7.36 plot the differences at each stenosis depth. At the 25 percent depth condition, for the two lowest blur conditions the differences are relatively small and constant. The correspondence between the standard deviations grows worse as the depth increases. At the 75 percent depth condition, there is a large amount of variation in the differences caused by both parameters.

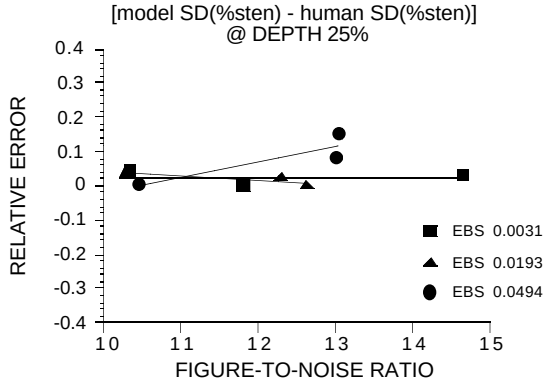


Figure 7.34 Model-human standard deviation differences as a function of EBS and FNR at stenosis depth 25%.

The Monte Carlo estimates for the model standard deviation predict least well trends in human standard deviations primarily at the highest blur and noise conditions (see for example the linear fits through EBS 0.0494 in Figure 7.33 or EBS 0.4444 in Figure 7.36). The model estimates alone (not shown) are substantially variable. EBS in particular has a great impact on model variability.

Sometimes, however, the model estimates become more variable as a function of increasing FNR (Figure 7.37). It is unclear why the certainty with which the core estimates were determined would improve in higher noise conditions. This phenomenon occurred in several places in the Monte Carlo results. It is this kind of behavior for the core-derived stenosis estimates that will have to be understood and modified before a Monte Carlo standard deviation could be used to predict human variability for this task. Section 7.6.3 discusses alternative and future measures of core model estimation and variability.

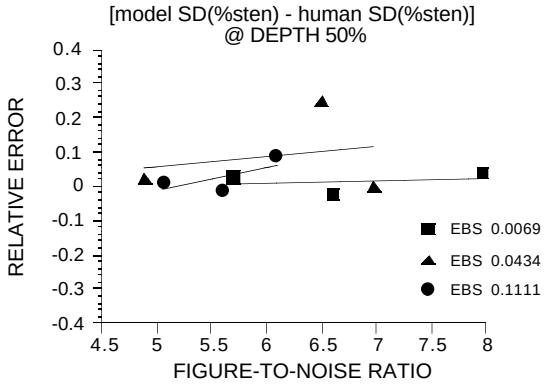


Figure 7.35. Model-human standard deviation differences as a function of EBS and FNR at stenosis depth 50%.

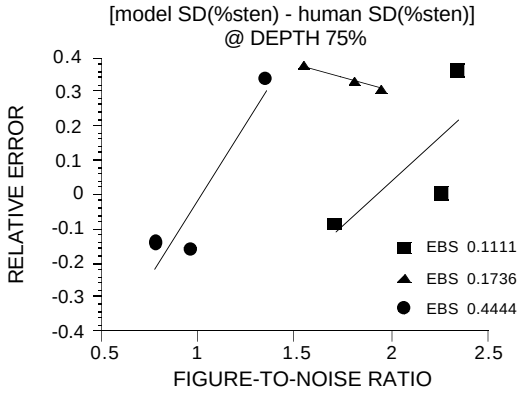


Figure 7.36. Model-human standard deviation differences as a function of EBS and FNR at stenosis depth 75%.

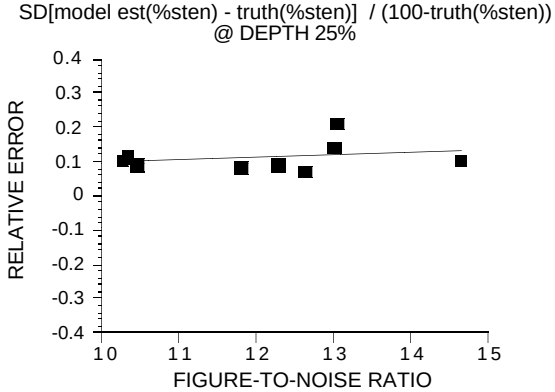


Figure 7.37. Monte Carlo standard deviations for stenosis estimation as a function of figure-to-noise ratio at stenosis depth 25%.

It was suggested in Chapter 2 that image quality can often be expressed using the mean and standard deviation in accuracy to form some kind of a "signal-to-noise ratio" describing performance. That is, both the mean accuracy and the standard deviation, or "certainty," are needed to fully characterize overall performance: a mean error with small variance is preferable to the same error with larger variance. Eventually, it would be desirable to achieve reliable model means and sensible, principled estimates of model variability that could be used in combination to characterize the quality of an image. As the best images are those for which observer mean error and standard deviation both approach zero, the figure-of-merit might not be a ratio but the square root of the sum of the squares of the two. The use of that form of a measure, however, is not feasible until mean performance of the model can be shown to parallel that of the human and until the model's variability can be understood and characterized more fully.

7.3.4 Standardized Differences

The differences between the model and human stenosis estimation relative errors were rarely greater than 0.10 at any given condition. An alternative way to look at the model performance is to ask whether it falls within the range of normal human variability. The model might be said to perform like a normal or average human if its mean, μ_{model}, falls with ζ standard deviations of the human mean, μ_{human}. Thus the measure is the difference between the means normalized by the standard deviation in the human mean:

$$\frac{\mu_{model} - \mu_{human}}{\sigma_{human}} < \zeta \qquad 7.2$$

If these standardized differences were always within some acceptable tolerance, such as 1, at all the experimental conditions, then the model might be thought of as no better or no worse than the average human from the experiment population. This is essentially the question that the statistical analyses ask, but it is informative to look at the data plotted in this way. Furthermore, the estimates for the human variability really should be estimates of confidence intervals that make a statement about the bounds of the normal population variance as opposed to the standard deviations for the observers in the experiment alone.

The plots below show these standardized differences as a function of stenosis depth (Figure 7.38), figure-to-noise ratio (Figure 7.39), and effective blur scale (Figure 7.40).

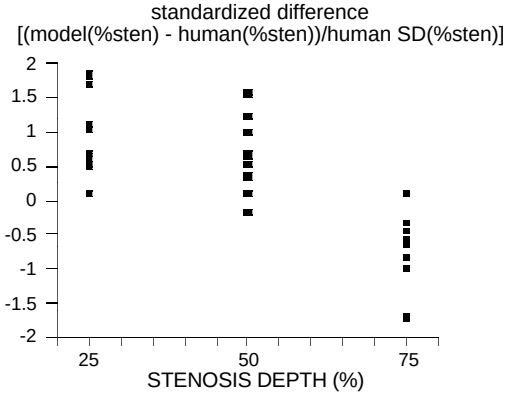


Figure 7.38. Standardized differences as a function of stenosis depth.

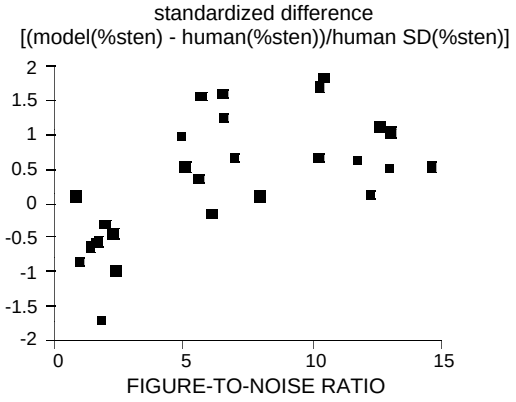


Figure 7.39. Standardized differences as a function of figure-to-noise ratio.

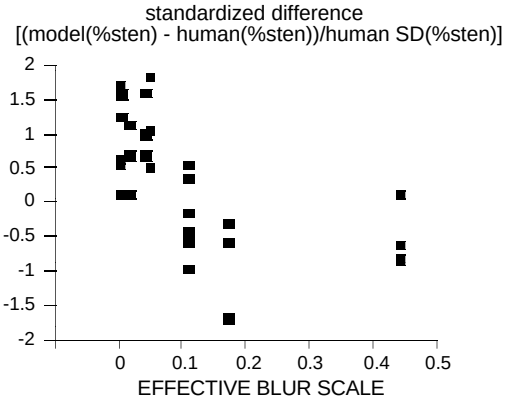


Figure 7.40. Standardized differences as a function of effective blur scale.

The standardized differences do in general increase as a function of decreasing FNR and increasing EBS much the way that the original results did. However, in many cases the standardized differences are less than 1, and they are never greater than 2. So in this sense, the model is reasonably predictive.

7.4 Model vs. Human Performance for Distance Estimation

Human and model estimates for the treatment field distance task are analyzed here. The clinical motive underlying this kind of an investigation is a desire to determine that image processing parameter combination which allows the most accurate estimation of the true physical status. In the application in this research, the different levels of SHAHE processing presumably cause systematic changes in field distance accuracy with respect to the true distance between the field and the vertebral anatomy.

The difference scores that are examined here are differences between model and human relative errors. Because this task involved the judgment of a range of treatment field clearance distances, it is important that the score assigned to the model and human be a relative error. The paramount question then is whether the relative errors made by the model parallel the human relative errors change in the same way as a function of the nine parameter conditions. The standard categorical ANOVA methods described in the introduction to this chapter were employed throughout the next two sections.

These relative error estimates of course require a value for the true distance between the two important edges involved in the task. Since no real truth is available for these images, a *designation* of truth was performed. The designations were made via a combination of computed methods and the author's judgment. The determinations were made on the preprocessed backgrounds. First, ridges of the magnitude of the derivative of the image intensities with respect to the horizontal, or x, direction, initiated by a starting position supplied by the author, were computed to determine the treatment field and vertebral body edges. Truth was taken to be the horizontal distance between the two ridges at a vertical position in the middle of the image. For eight of the 27 backgrounds, the edge positions determined in this manner did not correspond to the positions that should have been chosen according to the directions for the task and were subsequently adjusted slightly by the author to make them consistent with that specification. The mean truth determined by this method for each of the nine experiment conditions is shown below in Table 7.9.

	contrast 2	contrast 7	contrast 12
gain 1	123.39 (32.7)	112.06 (33.1)	119.72 (30.3)
gain 3	129.61 (31.1)	122.22 (33.4)	114.72 (35.6)
gain 5	123.11 (29.8)	112.33 (28.2)	118.89 (34.2)

Table 7.9. Means (and standard deviations) of the "true" treatment field clearance distances in the nine portal imaging experiment conditions.

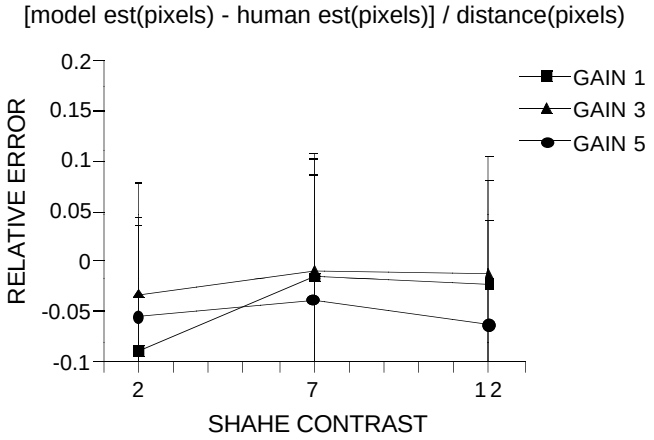
Clearly this method of determining truth is biased. Computed estimates of edge positions may produce truths that are more similar to the model's estimates. However, this author agreed with the positions determined by the derivative measurements or adjusted the results in the cases where the estimates were incorrect. Yet the author is biased as well. Thus, the values are not so much "truth" as rough estimates of how far apart in the original image the edges were. These values for truth made it possible to calculate relative errors as the outcome for these analyses. It also allowed inspection of trends with respect to some value, even if it was not *bona fide* truth, so that human and model accuracy could be examined alone.

Data in this section are reported as differences between the human and model mean relative errors for the 18 backgrounds at each condition. The units and levels of the SHAHE parameters are as discussed in Sections 5.1 and 6.3.

The source table for the ANOVA is shown in Table 7.10. The difference scores for the nine experimental conditions are shown in Figure 7.41. The scores are almost all negative, meaning that the model distance errors were in general smaller than those of the human. The sharp trend toward model underestimation relative to the human at the gain 1 and contrast 2 condition causes the significant interaction between the contrast and gain variables: the ordering of the scores with respect to gain is different at each contrast (F=6.87, G-G p=0.0021).

SOURCE	F	G-G Pr > F
gain	8.81	0.0050
contrast	6.78	0.0123
gain•contrast	6.87	0.0021

Table 7.10. Source table for the analysis of model-human distance estimation differences.



	contrast 2	contrast 7	contrast 12
gain 1	-0.090 (0.13)	-0.016 (0.12)	-0.023 (0.11)
gain 3	-0.032 (0.11)	-0.010 (0.11)	-0.055 (0.10)
gain 5	-0.055 (0.10)	-0.038 (0.12)	-0.063 (0.10)

Figure 7.41. Model-human distance estimation differences as a function of SHAHE gain and contrast.

The correspondence between the model and human results is roughly constant and near zero at the gain 3 condition. It is also fairly constant at the gain 5 condition. Interestingly, the greatest variation in the difference

scores occurs at the gain 1 condition, and the greatest disparity across all nine conditions occurs at the gain 1 and contrast 2 condition. It is these conditions that may be considered "mildly" enhanced or most similar to the original images.

The significant interaction of the contrast and gain parameters for the model-human relative error differences does not permit examination of the main effects of those parameters. Figure 7.41 demonstrates that the interaction is not simply ordinal, and thus averaging scores across the gain or contrast conditions would be meaningless and perhaps misleading.

There are fundamental differences in the comparability of the model and human results when the SHAHE contrast and gain parameters were adjusted. The significant two-way interaction suggests that the relative errors from the model were not able to predict those from the human for the range of SHAHE parameters studied here.

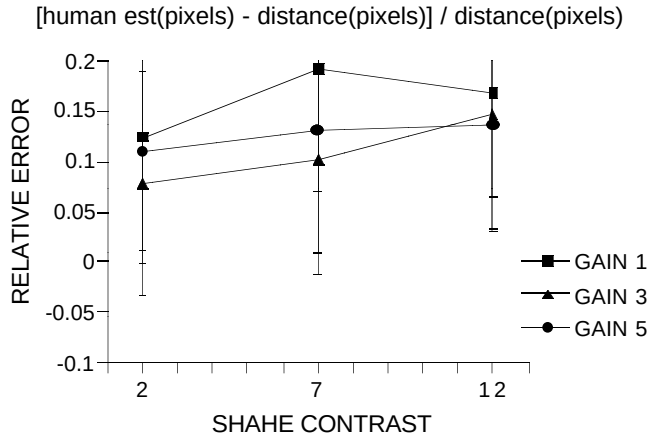
7.5 Human and Model Performance for Distance Estimation

What the previous analysis can not determine is the accuracy of the human or model alone. In this section, the analysis compares model and human estimates against the "true" edge-to-edge distance for each patient case. There were several reasons for these additional analyses. First, estimation trends, such as a systematic increase or decrease in perceived distance as caused by adjusting one of the parameters, could provide some feedback to the current users of SHAHE about what parameter settings might be optimal. Second, it would be particularly encouraging to know if there are conditions where the model paralleled human accuracy when human accuracy was not simply constant but changed a great deal as a function of the parameters. Third, where there were discrepancies between the model and human results, it would be of interest to know whether it was the model or the human that best estimated the truth. Finally, there was no way to know whether the humans used the appropriate edges in each image in making their judgments. Some conclusions can be drawn in this regard by examining the inter-observer variability in the human accuracy data.

7.5.1 Human Accuracy

The main effects and interactions for the human distance accuracy estimates are shown in Table 7.11.

Figure 7.42 depicts the relative errors in human estimation with respect to the determination of truth. Humans exclusively overestimated the true distance, and relative errors were as great as 0.192 in one of the conditions. The errors seem to increase in general with an increase in the contrast parameter. However, the data at each contrast level are not ordered with respect to the gain parameter: it is the gain 3 condition at which performance is best. The statistical analysis concluded that there was a significant interaction between the two independent variables (F=6.36, G-G p=0.0030).



	contrast 2	contrast 7	contrast 12
gain 1	0.124 (0.13)	0.192 (0.12)	0.167 (0.10)
gain 3	0.078 (0.11)	0.101 (0.11)	0.146 (0.12)
gain 5	0.110 (0.10)	0.131 (0.12)	0.135 (0.10)

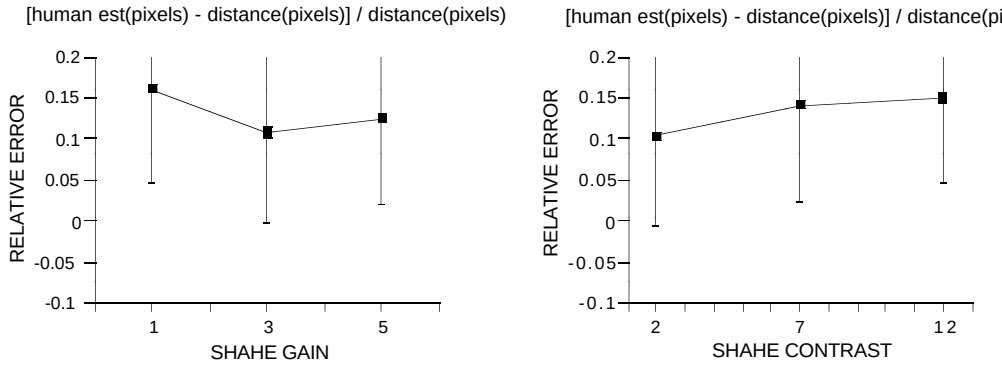
Figure 7.42. Human distance estimation errors as a function of SHAHE gain and contrast.

While there was an interaction of the two independent variables, the relationship is sufficiently ordinal that it is worth looking at the main effects alone. The plots (Figure 7.43) confirm the overall trends for the two parameters. Apparently some intermediate amount of edge sharpening was beneficial to the humans in

SOURCE	F	G-G Pr > F
gain	20.34	0.0001
contrast	10.96	0.0022
gain•contrast	6.36	0.0030

Table 7.11. Source table for the analysis of human distance estimation errors.

determining the true distance. Conversely, a systematic increase in contrast alone only resulted in poorer accuracy.



gain 1	gain 3	gain 5
0.161 (0.11)	0.108 (0.11)	0.126 (0.10)

contrast 2	contrast 7	contrast 12
0.104 (0.11)	0.141 (0.12)	0.150 (0.10)

Figure 7.43. Human distance estimation errors for the main effects of gain (left) and contrast (right).

A separate statistical analysis was performed on the human accuracy data to test for the presence of a "practice" effect. The analysis tested whether there were significant differences between overall relative error at the first and last (eighteenth) trial in every condition. While the mean accuracy worsened slightly, the analysis indicated that this was not significant. There was no significant main effect ($F=0.51$, $p=0.4901$), nor were there other interactions involving the trial variable.

7.5.1 Model Accuracy

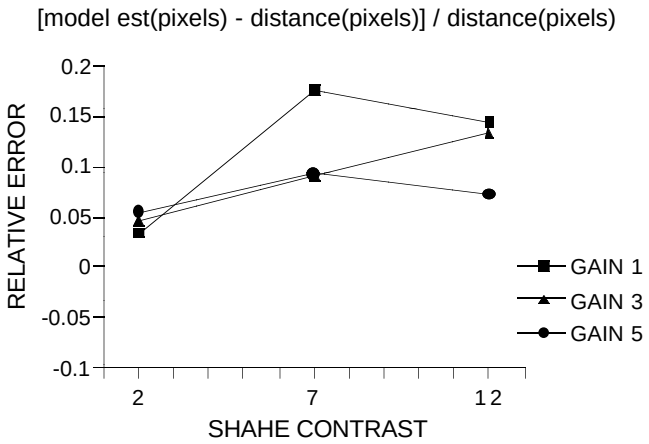
As with the angiography data, no statistical analyses were attempted for the model data alone. However, the trends in model errors are quite understandable.

Figure 7.44 shows the model accuracy data for all nine experimental conditions. Like the results from the human observers, model estimates are all greater than the true distance. At the contrast 2 setting, model accuracy is good and similar for all three gain values. As the contrast is increased, it seems the gain parameter must be increased as well in order to achieve the same accuracy.

The good accuracy at the condition which could be considered least enhanced (contrast 2, gain 1) represents a sharp improvement over the model's accuracy at the other two contrast levels for the gain 1 parameter. It is the model's particularly accurate performance that causes poor correspondence with the human results at this condition.

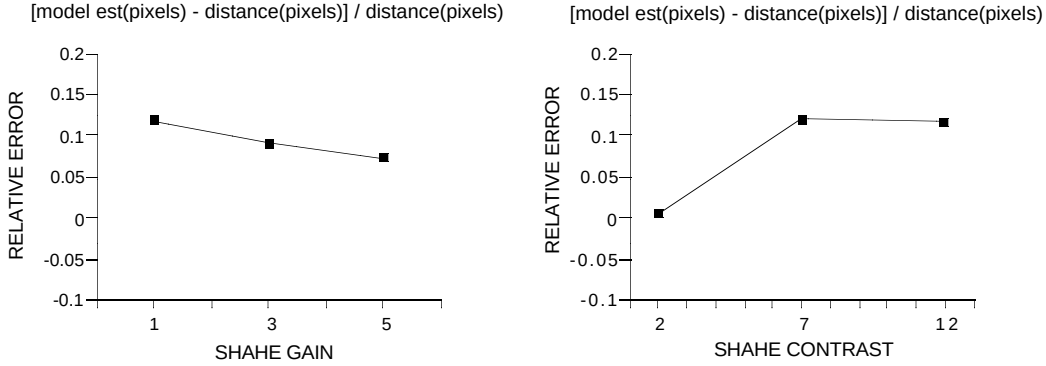
The main effects for gain and contrast (Figure 7.45) demonstrate clearly the trends in model accuracy. Accuracy increases nearly linearly with increasing gain. Conversely, accuracy decreases as the SHAHE contrast parameter is increased.

These plots suggest that the higher gain settings are helping to counteract the distance-increasing artifact that appears when increasing the contrast parameter. Unfortunately, the human results can not be accounted for by this explanation. While the model behavior seems to make sense in terms of this trade-off of the effects of the parameters, somehow human behavior did not also exhibit this effect. At a given contrast level, it was always the gain 3 setting at which humans performed most accurately. Ultimately it is the human performance that must be predicted by the model.



	contrast 2	contrast 7	contrast 12
gain 1	0.034	0.176	0.144
gain 3	0.046	0.091	0.134
gain 5	0.055	0.093	0.072

Figure 7.44. Model distance estimation errors as a function of SHAHE gain and contrast.



gain 1	gain 3	gain 5
0.118	0.091	0.073

contrast 2	contrast 7	contrast 12
0.049	0.120	0.117

Figure 7.45. Model distance estimation errors for the main effects of gain (left) and contrast (right).

7.6 Discussion

There is now much that has been shown about the correspondence between these implementations of the core model and the human results. In most cases the agreement was not sufficient to allow the model to be used presently to predict medical image quality. Yet there is nothing about the theory or results that suggests that this approach does not have potential for that purpose. The task remaining is to speculate about the source of the differences and point out what was learned so that future investigations and visual model development may capitalize on these endeavors.

7.6.1 Experimental Design Modifications

There are several issues common to the designs of both experiments that deserve reflection in light of the analyses in this chapter. First, of course it is always advantageous to collect data from many human observers; the results of the human observer experiments in this research could only have been solidified with more data. That there were significant effects for both human accuracy and model-human differences, however, suggests that there were enough observers. Statistical power calculations that make a determination of the number of observers needed to demonstrate an effect would have been necessary only in the event of *non*-significant results. Second, the experimental designs should have been counterbalanced by "block" so that a trial from every experimental

condition was presented before moving on to another such block. The random ordering used in these studies is not as effective in designing against inevitable practice and carryover effects. Third, it also would have been useful to obtain a measure of intra-observer variability. The two experiments never repeated any of the images in the experiment. That kind of a measure would have said much about the certainty and consistency with which these relatively untrained observers performed the tasks. Fourth, the experimental designs both utilized different backgrounds at each condition. This allowed the whole experiment to contain a broad sample of potential cases. However, the means and variances in each condition were different from each other in part because of the different backgrounds as well as because of the effects of the variables of interest. It might have been better to use the same backgrounds in every condition and suffer a lack of generalizability at the benefit of decreased variance. All these considerations are not to say that the reader ought to have serious doubts about the validity of the human results. The designs and experiments were developed with care, and the resulting trends in human accuracy are mostly understandable. Nevertheless, these issues related to believability in the gold standard human data ought to be mentioned here and heeded in future investigations. The compromises in this experimental design, clinical applicability and a reasonable experiment duration at the expense of psychophysical control and statistical power, are typical of the demands of an engineering application.

Both experiments utilized measuring tools for indicating the task estimates. The tools made the tasks in some respects difficult and may have contributed additional variance to the human results. The problems incurred with the measuring tool were more acute in the portal image distance estimation task. To perform that task, observers had to judge the distance between two edges that did not belong to the same object and that were well apart from each other. To maintain and translate that visual representation to the measuring tool several degrees of visual angle to the right was perceptually difficult and no doubt fraught with error.

Chapter 6 provided the rationale for why the tools were necessary and implemented in the way that they were. There is little recourse from these particular decisions. The visual system has at its disposal a vast array of mechanisms and strategies for carrying out a particular task. The visual model that was tested here is a model only for shape representation, and the medical image tasks for which the core model hypothesizes mechanisms are those that involve the judgment of shape. The measuring tools in both experiments were intended to make it such that the visual operations performed by the human observers were comparable to those that the core model was developed to mimic. Otherwise the core implementation would have been tested against perceptual strategies that it does not purport to model.

At each of the conditions in both experiments, there are fewer observations for the model than from the combination of the human data. The human data points were means for thirteen observers while the model is in effect a single observer. Those increased observations probably helped produce more stable human results. However, the Monte Carlo simulations, which determined a mean model performance over many noise realizations for the images in the experiments, were not any more successful at predicting human performance. Alternatively, the model could have been used to easily compute estimates for many more and different images than those that were used in the human observer experiments. This would have provided a better assessment of the model's behavior for these blur and noise conditions for a larger range of potential image backgrounds. However, it was only fair to compare model performance for images for which there was human data. In the end, the hypothesis for the use of this approach is that the model alone performs on any given image in a way that parallels overall human performance.

Part of what makes the angiography experiment and its analyses difficult to interpret is that the three parameters, stenosis depth, EBS, and FNR, are potentially tangled. The principles of the human visual system specify that the accuracy with which an observer can judge the depth of a stenosis will be roughly proportional to the depth of the stenosis. Trends in stenosis estimation might be expected on that basis alone. At the same time, the effect of quantities like blur and noise on perceivability and interpretability are dependent on the size of the objects in question. The analysis and interpretation are simpler, have more power, and are easier to interpret, when the parameters in the design define an orthogonal relationship. The conclusion to Chapter 6 discussed how images for this experiment could have been produced so as to allow an orthogonal design with FNR and EBS as the parameters. Stenosis depth was a parameter in this experiment because it was important to test the predictability of the model for the range of potential vessel constrictions that might be encountered in typical clinical practice. Since the EBS and FNR parameters reflect the vessel width information, it might be preferable in a subsequent experiment to not study stenosis depth but instead vary it in a random fashion the way that background and vessel path were. That reduces the interpretation of the results to examining only the physical property variables of interest.

The requirement that the differences between the model and human be constant across the experiment conditions may perhaps be too stringent. If the relationship between the model and human errors were known to be monotonic, that is, that the differences between the two were monotonically divergent or convergent, or if the relationship satisfied other properties that guaranteed the same optimum as a function of the parameters, then the model might still be used to effectively predict best human performance. The analysis in that case consists of testing whether human performance as a function of the independent variables of interest is predicted by model

performance that is allowed to possess a possibly different slope. These alternative, exploratory analyses unfortunately must be left for future investigations.

It is inevitable in any experimental investigation, particularly one that utilizes a novel design and tests a novel hypothesis, that factors in the experimental design and analysis together with issues from the interpretation of the results will stimulate subsequent studies. It is rare that in a single attempt a flawless design is developed that possesses a corresponding analysis that is elegantly appropriate. Two different experiments were conducted in this research in order to explore the feasibility and predictability of this model-based medical image quality approach in more than just a single imaging modality. It is hypothesized that eventually this kind of an approach could be used for many tasks with any imaging system. Yet the present analysis suggests many modifications for further experimentation. The angiography experiment, owing to recent theoretical considerations that affected the way that the parameters in that experiment were quantified, possessed an experimental design that required in turn a complex statistical analysis. And in both experiments, the feedback about both the design as well as the model under investigation can be used in a subsequent study. What might have been a more reasonable goal for this dissertation is to conduct a short series of successively refined experiments that hone the experimental techniques and at the same time work toward a conclusive demonstration of the model's usefulness.

7.6.2 Angiography Experiment Overview

That the absolute errors in relative accuracy between model and human accuracy for stenosis estimation were rarely more than 0.10, an absolute error of 10 percent in the percent stenosis measure, is indeed encouraging. The model performed accurately and in many cases within the bounds of normal human variability. Furthermore, the model was often influenced predictably in an overall manner by the variation in effective blur and figure-to-noise properties. Most importantly, the variation in the accuracy of both model and human estimates that occurred as a function of the extent of effective blur or figure-to-noise ratio was in an overall sense very similar. However, what was not achieved was perfect parallelism across specific parameter combinations. The model results simply did not track the human data closely enough in the individual conditions to guarantee that the model would be generally useful as a method for localizing maximal human accuracy.

To begin with, there were fundamental differences between the model and human results for the basic task of stenosis estimation for plain vessels. Humans overestimated the extent of plain curved vessels at the 25 percent depth condition while the model exhibited rather constant accuracy. The way that model and human stenosis estimation were in turn influenced by the imposition of the experimental conditions was different: the human data were simply shifted in the direction of underestimation while the model moved toward overestimation at mild constrictions and underestimation at severe constrictions. It was true that the trends in accuracy with stenosis depth were the same for both: the model and human data both revealed increasing underestimation for increasing stenosis constriction. However, the changes in the estimates from plain to embedded vessels for the model as opposed to the humans is an important indication that the model did not respond to noise and blur in the same way that humans did.

At any given stenosis depth condition, the model's behavior as a function of noise and blur was often difficult to summarize. Sometimes, model stenosis estimation was relatively unaffected, particularly by noise (Figure 7.16, right, for instance). When this was the case, it was not entirely surprising: the core representation can be relatively tolerant of smaller scale noise and blur. In other instances however, the model's estimates are quite variable (Figure 7.17). Unfortunately, sometimes these trends in model accuracy are not predictable. For a fixed EBS, accuracy as a function of increasing FNR might decrease, for instance. Finally, Figures 7.22 through 7.24 show the largest variations in the difference scores at each stenosis depth occur at the largest EBS and smallest FNR condition. A generalization of the results then is that the model's stenosis estimates for blur and noise were least able to predict human accuracy in the most degraded conditions.

It was important for the purposes of these experiments to use a range of physical conditions that was broad enough that human and model performance would vary as a function of those conditions. As a result, the highest EBS and lowest FNR conditions in the angiography experiment represented very noisy and blurred images. In fact, the images were probably degraded more so than what might be encountered in clinical practice. That the model estimates were least predictive in these conditions is not as discouraging when the extremity of these noise and blur degradations is taken into account.

7.6.3 Monte Carlo Methods for Further Characterization

The Monte Carlo simulations that were presented in Section 7.3.3 were simply an attempt at developing some form of a variability estimate that could be ascribed to the model that might be comparable to an inter- and intra- observer standard deviation. Standard deviations, over many instantiations of noise, in the percent stenosis estimates (as computed by the protocol described in Section 4.4 for determining a stenotic and normal width from the core) were used as the basis for the comparison with the human deviations. Alternatively, related variability estimates could just as easily have been computed from, for example, the standard deviation in the estimate of the normal or stenotic width of the vessel alone or even in the positional shift in the calculated core center. Moreover,

there are theoretical means of determining the certainty in core scale and position that relate the amplitude and periodicity in the medialness of a noise distribution to the amplitude and curvature of the medialness peak at a figure center. More work must be done to develop and understand these certainty and stability aspects of the core's behavior. When that is done, these estimates in core variability may be reasonable predictors of the certainty with which humans can perform this and other estimation tasks.

The Monte Carlo simulations that were done for this study represent only a first step in the process of understanding the characteristics of the core model's operation under realistic image conditions. Neither the means nor the standard deviations produced by the simulation were statistically successful in predicting human performance. However, Monte Carlo methods should now be used to chart with finer sampling and broader range the model's estimation behavior for blur and noise. Also, in addition to the blur and noise manipulations, it will be important to study and characterize core model estimation for realistic variability over *shape change*, such as normal vessel width, stenosis properties such as asymmetry and axial length, and vessel boundary texture. Variability in those characteristics maybe even more representative of clinical variation than variability over noise instantiations. It is this kind of a simulation approach that can now be used to chart the model's performance over a broad range of potential conditions to characterize and adjust the operation of the model.

7.6.4 Perceptual Measures of Blur and Noise

This dissertation utilized two relatively novel measures for describing the extent of blur and noise in the angiographic images. The effective blur scale (EBS) and figure-to-noise ratio (FNR) were intended to quantify the degradations in perceptual units. They are reasonable guesses about how to measure those quantities that were based on the zoom invariance of the visual system that has been established to be a good approximation. It is of course possible that either or both the EBS or FNR measures used here are not appropriate or entirely valid. For example, both measures incorporated only stenotic width. The actual stenosis estimation task involved the judgment of two widths, the stenotic width and a normal width. That normal width probably ought to enter into the formulas for both measures. On the other hand, observers could have made the normal judgment at any position(s) along the vessel that were more or less noisy, so it is unclear how the extent of noise at the normal width should be characterized. Also, there is psychophysical evidence (Section 3.2) as well as data from these experiments (Figure 7.11) that suggest that the human visual system is not perfectly zoom invariant. It may be the case that the experimental results that do describe the form of the zoom invariance relationship should be used instead of just a figural width estimate alone in the EBS and FNR formulas. Lastly, the effective stenosis width is something slightly greater than the zero-scale width values that were specified here. The perceptual blur imposed by the visual system will cause some amount of widening of the vessels. That effect is likely very small and difficult to quantify. Nonetheless, it is worth considering whether the experimental results from this research might have something to say about an "effective depth" that could in turn be used in the measures of EBS, FNR, and even relative error. All these considerations only adds emphasis to the fact that the perceptual measures of blur and noise used here may not tell the entire story.

There is however good evidence that the EBS and FNR parameters were perhaps a better means of characterizing the effects of blur and noise than were the original parameters. Human accuracy data is plotted below as a function of the three Gaussian spatial widths (σ) that were used to blur the experiment images (Figure 7.46). Similarly, the same data is plotted as a function of the original noise parameter (Figure 7.47). The three values for the original noise parameter are the maximum intensities to which the experiment images were scaled prior to addition of Poisson noise (see Section 4.3.4 and 6.1).

In both the noise and blur cases, accuracy at any single setting appears to worsen as the depth parameter increases. Accuracy at the 25 percent depth condition in particular appears to be well segregated from the accuracy at the other two depths. Furthermore, the range of accuracies at any blur or noise level is fairly broad. If "noise 200," for instance, were a good characterization of some amount of noise, one would expect similar accuracy at all the conditions that were described by that level. Instead, the accuracies segregate by width. Finally, the linear fits through the data from the 27 conditions that are shown as a function of original blur and noise parameters in Figure 7.48 show that in an overall sense those parameters just barely capture the intended effect of variation in errors. Noise and blur relationships like those in Figures 7.5 and 7.6 seem more appropriate. Relative errors changed more and in a roughly linear fashion when the EBS or FNR parameters were varied.

The desired blur or noise characterization has the property that it has a linear relationship to performance: equal increments in estimation accuracy are the result of equal increments (or decrements) in the perceptual descriptor. The perfect definitions for the parameters in the angiography experiment would cause measured accuracy to form a hyperplane as a function of the three parameters. It would be difficult to depict this four-dimensional data set, and it is unlikely that the EBS and FNR parameters or the experimental data form such a perfect relationship anyway. However, it does appear from the previous argument that EBS and FNR were a reasonable step toward developing an appropriate measure.

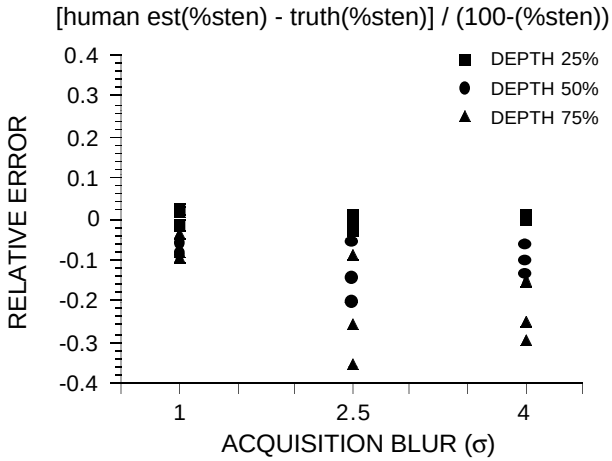


Figure 7.46. Human stenosis estimation errors as a function of the three Gaussian blur scales that were used to simulate the blur of the acquisition system in the experiment images. The three stenosis depths are also labeled according to the legend.

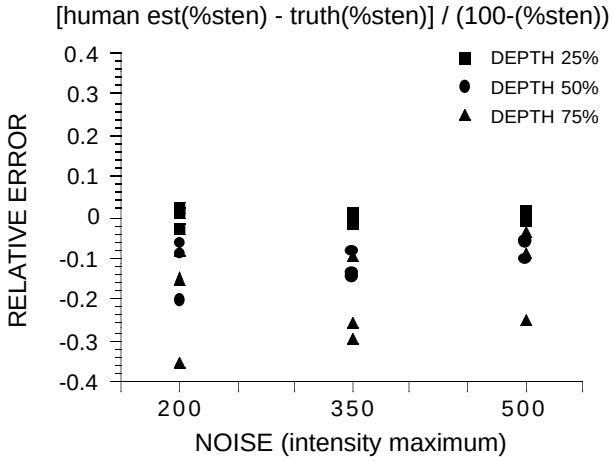


Figure 7.47. Human stenosis estimation errors as a function of the three intensity maxima used in scaling the experiment images prior to the addition of Poisson noise. The three stenosis depths are also labeled.

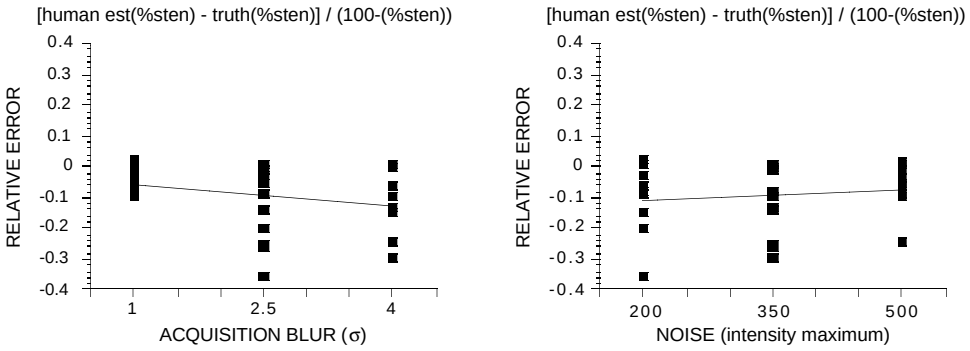


Figure 7.48. Human stenosis estimation errors as a function of the original blur (left) and noise (right) levels that were used to generate the experiment images. Linear fits are shown through all 27 experiment errors.

These questions beg a simple psychophysical study that could measure figure width estimation accuracy in the presence of blur and noise. It is likely that the correct descriptors could be derived *post-hoc* from the data.

Furthermore, if the core model truly is representative of the manner in which humans represent figures, then the way that the perceptual effects of blur and noise ought to be quantified is by describing their effects on medialness. As mentioned in the previous section, there may be ways to characterize the amplitude and curvature of a figure's medialness distribution such that it is those measures that best represent the impact of the blur and noise relevant to a particular figure.

7.6.5 Portal Imaging Experiment Overview

The model performance for the portal imaging distance estimation task was both quite good and sensible. The model performed with a maximum relative error over all nine conditions of 0.176. Furthermore, the trends in the model results seem to make sense in the perspective of the proposed purpose of SHAHE. That is, it appeared that an increase in the gain parameter to the unsharp masking preprocess was necessary to counteract the decrease in accuracy that came about for the higher settings of the contrast parameter. This is entirely consistent with the motivation for SHAHE that was put forth in Section 5.1.

While the accuracy of the model was on the order of, and in fact slightly better than, that of the humans, the changes in human accuracy as a function of the SHAHE parameters were not sufficiently paralleled by the variations in the model performance. There were significant main effects and an interaction for the differences scores as a function of both parameters. The human results indicate that the intermediate gain setting produced the best accuracy regardless of the contrast level. It is possible that the gain parameter has an influence on human perception of features or properties that are not captured under the core representation.

The portal imaging experiment did not possess so many of the design and analysis difficulties as did the angiography experiment. However there were more problems with the human performance of the task. The observers by their own admission had trouble performing the distance estimation task with the measuring tool. And while there was no intra-observer variability estimate, the inter-observer standard deviations that are plotted for that data indicate that the observers were at least performing the task differently. One useful addition might have been some sort of rough markers, brackets at the top of the image for instance, to indicate the general position of the edges to be used. Furthermore, a simple experiment, akin to the plain vessel angiography experiment, could be carried to measure the baseline variability that arose from using the measuring tool. Stimuli like lines and bars, that would eliminate any uncertainty as to how to locate the edges involved in the estimate, could be judged with the measuring tool. In this way, the variability that potentially resulted from the use of the wrong anatomical edges could be estimated to be the variability beyond that attributed to the performance of the task with the measuring tool.

The comparison of the model and human distance estimates with the rough designation of truth allowed several important conclusions. First, best human performance occurred for the lowest contrast setting and intermediate level of gain. This suggests that the optimal application of SHAHE is as a "mild" enhancement. Accuracy for unenhanced images was not measured in this experiment, but experiments by others have shown that some amount of SHAHE processing is beneficial. But apparently the increase in edge artifacts and noise that accompanies the increase in the contrast parameter did negatively affect human accuracy at least for the task, images, and parameter values used in this experiment. This observation is helpful but not surprising feedback for SHAHE users. The next paragraph discusses further the use of unenhanced images in this study. Second, as evidenced by the error bars in Figures 7.42 and 7.43, interobserver variability for this difficult distance estimation task was quite high. It appears that each of the observers was performing the task differently. That is not to say that each observer was not internally consistent or was not influenced in a predictable and consistent way by the experimental conditions. It is difficult to know whether some observers were using the wrong edges in their decisions. Everything short of having a preceptor standing over the observer who would verify every decision, which in itself might have been susceptible to bias, was done to ensure that observers utilized the intended edges for the task. Finally, the truth determinations allow the conclusion that the model results are in general only slightly more accurate than those of the human. Thus it cannot be said that somehow large disparities in overall accuracy are the source of particular discrepancies.

It might have been useful, for the purposes of the evaluation of the SHAHE processing technique, to include unprocessed images in the experiment. It would have been interesting to know whether human (and model) accuracy for the task was significantly better at any of the SHAHE settings than for the unprocessed images. However, the unprocessed images were difficult for untrained readers to interpret: the edges involved in the task were often nearly impossible to locate because of the poor contrast. The human observers used in this experiment might have unreliably localized the edges in these more difficult images. Interestingly, the computed means of determining the true distance in the unprocessed images, which utilized essentially the same analysis methods as does the core model, rarely had difficulty in landing upon the correct edges.

7.6.6 Computational Modifications

There are multiple opportunities for modifying the core model implementations for potentially enhanced predictability with respect to the human task estimates. There were decisions that were made in this research for

several stages of the model computations: a medialness operator and ridge tracking mechanism were chosen, and protocols were developed for producing an estimate from the core.

A number of medialness kernels might have been chosen for this research. The Laplacian of a Gaussian medialness kernel was chosen for the angiography study because of its demonstrated robustness in the presence of the blur and noise degradations that were present in that experiment. The Laplacian kernel has a large region of integration at its center that, for a simple uniform figure like a blood vessel, has a way of averaging out the noise and tracking powerfully that figure middle. However there are several reasons why this Laplacian operator might not be as likely to be a mechanism utilized by the human visual system. First, in order to represent figures that may have distracters intervening between the figure boundaries or shading variations across the figure, the visual system probably *links* boundary information to figure middles. That is, the range of interaction at the figure boundary, while increasing with figural width, is still a fraction of that width. It is unlikely that a perceptual operator, like the Laplacian, with such a significant component at its center could be the mechanism for representing those kinds of figures. Second, it is physiologically implausible that the many synaptic connections that would be needed to carry out the measurements that the Laplacian represents would exist in the visual system. Instead, the visual system is likely to utilize a rather limited number of boundariness cells whose linkage is established and strengthened by excitatory feedback. It is the medialness kernels that mimic the linkage of a few important, prominent boundary locations at a proportional distance from the kernel center that are most plausible. So there is opportunity for research into justifying particular medialness kernels and in turn using them for the angiography task in this research.

According to these arguments, the medialness kernel used in the portal imaging experiment seems very sensible. That kernel represents a linkage of only two special boundariness locations that are along the horizontal. The idea is that the knowledge of the task would induce excitatory feedback to the boundariness receptors at the needed, known positions. The problem in devising that kernel is to know how to choose the proportionality constant that specifies the relationship between figure half-width and boundariness scale. For these experiments, that value (8) was chosen so that the core calculations were computationally robust. For the typical treatment field clearance distances in question, that proportionality constant resulted in boundariness scales that did not key on minute structures but was able with good precision to locate the edges of interest. It appears from the data as though the model was able to locate the edges too well. Model performance in all of the SHAHE parameter conditions was more accurate than that for humans. For that distance computation to produce more predictable estimates, it will be necessary to set that proportionality constant as dictated by whatever psychophysical evidence can be brought to bear.

The particular ridge extraction technique used in this work was at the time of the research the most stable and efficient method available. It was most likely to produce from the underlying medialness data set a continuous core that met the criteria for extracting the task estimates. Under the most degraded noise and blur conditions in the angiography experiment, however, core finding and tracking was more likely to fail. Preparation for the experiments included a scheme for producing a set of images for which the core calculations were entirely successful. Many background/vessel combinations were thrown away before a complete set of images was generated. The core calculations that were tested against the human results were thus from a non-random subset of all the images that might have been used. The correspondence between the model and the human was poorest in the highest blur and noise conditions. Yet this may not be a legitimate conclusion regarding the model when the results were computed in conditions where the model normally had trouble computing an estimate.

Ridge following ceases (or can not be initiated) when the magnitude of the noise and blur causes the medialness to take on a distribution where the conditions for the ridge definition are not met. Blurring of objects makes the resulting medialness less sharp, or "flattened," by an amount proportional to the scale of the blurring. In that case, the eigenvalues that are the basis for the ridge decisions may become zero. The ridge criteria are evaluated with respect to directions of maximally negative second derivatives; where the derivative is locally zero, the ridges do not exist or ridge tracking must terminate. Noise introduces into the medialness quasi-periodic perturbations whose magnitudes are related to the severity of the noise. Cores may result that can pass nearby in scale space to the core of interest. While generically it is known that cores do not branch, with a finite numerical representation the eigensystem solving may encounter the conditions for a branch when these "confusing" ridges are nearby.

There are numerous computational advances that have been proposed for improving core stability in the presence of blur and noise. Several of these improvements aim to make core construction operate more like the visual system, which is based on neural excitations and inhibitions. In particular, Morse, in his dissertation,[2] has proposed a set of feedback mechanisms that allow the cooperation of medialness and ridge formation. First, a credit attribution scheme allows iterative refinement of medialness by strengthening the connections between boundary receptors and hypothesized medialness cells for those medialness cells which were highly activated after

[2]B.S. Morse, "Computation of Object Cores from Grey-Level Images," Ph.D. dissertation, (University of North Carolina-Chapel Hill, 1994).

a previous iteration. This process alone serves to refine or sharpen the medialness to in turn improve the success of subsequent ridge tracking. Second, the presence of ridges that are found can feedback to medialness formation. Inhibition along the ridge in the cross-ridge directions can in effect sharpen the medialness in the region of the ridge. Excitation of the medialness in scale space positions tangent to the ridge at its terminus can allow ridges separated by smaller gaps of weaker, flatter medialness to be connected up.

Eberly has recently proposed improvements to the "flow" process whereby ridges are found and tracked.[3] Ridges are zeros of the combined function $(P^2 + M_\sigma^2)/2$. $P^2 = \vec{v} \bullet Df$ is the derivative in the direction of maximal second derivative, and M_σ is the derivative in the scale dimension. A path from the initial guess to a point on the ridge can be determined by gradient descent or conjugate gradient methods. Convergence is not guaranteed for these methods, and they require eigensystem solutions that are susceptible to numerical error. Eberly is developing a bisection method that bounds the root that is the ridge and should consistently locate a ridge from the initial guess. Often the core computation failures in the presence of blur and noise were the result of failures to even initiate core traversal with the initial guess.

There is little evidence for how information might be extracted and combined within or along subsections of a single core to perform specialized judgments like stenosis or gap object estimation. The protocols for the extraction of the pertinent information from the core were grounded primarily in common sense. There are many adjustments to these protocols that could be made. The angiography estimation consisted of several arbitrary decisions about, for example, sliding window width and where to position the sliding window. The distance estimation used a simple mean in an arbitrarily sized window along the core. These aspects of the adaptation of the core model are perhaps the weakest components of the model computations of these tasks.

7.6.7 Final Remarks

This initial investigation has shown the two implementations of the core model that were studied in this research to not be tightly correlated with the human estimates. Particularly for the angiography conditions, it will be important in the future to establish a testable behavior for the model. It will no longer be useful to simply generate results for the model and test whether they are predictive of human data. Research into the core model and whatever implementation of it are used to perform these tasks must advance to the point where the model can be shown to exhibit predictable, understandable trends in accuracy or variability as a function of these physical properties. It is only then that a test of the model makes sense. Furthermore, for stenosis estimation, model performance of the task must first and foremost be made to parallel the human results for plain vessels before the model has any hope of being useful in sorting out the effects of noise and blur on that task. So until the model demonstrates a sensible behavior, research must be invested in tuning it to have one. It plainly *was* the case in the portal imaging study that an explainable behavior was exhibited by the model. Unfortunately, those results were perhaps more believable (or at least understandable) than the human results. But in that case the research can take a different direction; efforts can henceforth be invested in understanding the human behavior and the nature of the discrepancies between the two.

The implementations of the core model that were adopted or invented for the purpose of carrying out the specific tasks in this research must be distinguished from the core model itself. The core model posits a set of mechanisms for the representation of a single figure. That construction is performed in a manner that is thought to be consistent with the operation of the human visual system, and some of the manifestations of these principles have been verified in psychophysical experiments. However, this discussion has mentioned how the choice of a medialness operator was probably not optimal for predicting human performance. The ridge computations are implementation decisions that are not specified by the core model itself, yet they must be robust and accurate if the medialness maximum that is the core is to be determined consistently in the presence of potential variations in physical characteristics of the image. Again, the extraction of information from the core for the performance is independent of the core model predictions. The result is that the insufficient predictability of many of these experimental results says little about the efficacy of the core model theory. There are a multitude of modifications that could be made based on this promising visual model that could be used in future studies of this image quality approach.

[3] D. Eberly, personal communication.