Evaluation of Supra-Threshold Perceptual Metrics for 3D Models

Ioan Cleju*
University of Konstanz

Dietmar Saupe†
University of Konstanz

Abstract

Measures of dissimilarity of 3D models are necessary in a wide range of applications such as geometry compression, simplification, and 3D model retrieval. In many cases a metric that models perceptual dissimilarity is desirable. Recently, metrics for 3D models have been evaluated in that respect using concepts such as just noticeable differences, rankings, and others. We propose a simple experimental setup for evaluating supra-threshold perception of 3D models in which users select models at equal perceptual distance to given pairs of models. We discuss the advantages of our approach and report the results of a field study comparing six objective distance measures applied to palettes of simplified reference models. We found that the objective measures are biased, and generally image-based metrics perform better than metrics based on the original 3D geometry.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism

Keywords: supra-threshold perception, perceptual metric, image-based metric, level-of-detail

1 Introduction

The impact of simplification and compression algorithms on human perception of 3D models can be analyzed by perceptual metrics. The de-facto standard method for error evaluation is the Hausdorff geometric distance, given by the Metro tool [Cignoni et al. 1998]. This measure constitutes a guarantee on the maximum geometric distance rather than reflecting the perceived distance between the models.

Perceptual distances between 3D models were mostly applied in the context of simplification and Level-Of-Detail (LOD) management. LOD management algorithms seek to minimize the computational effort, while maintaining the rendering quality. [Luebke et al. 2002] provides a good overview on LOD techniques and the relation to perception. Most approaches to perceptual distances apply the Just Noticeable Difference (JND) concept to 2D renderings of the models [Reddy 1997; Reddy 2001; Luebke and Hallen 2001] or directly to 3D models [Cheng and Boulanger 2005]. A commonly used framework for constructing perceptual metrics uses the Contrast Sensitivity Function [Luebke et al. 2002] to model the human vision system. As a result, the distances suffer of the supra-threshold problem, being accurate only if the difference between the models is at the limit of perceivability [Wang et al. 2004; Lindstrom 2000].

Experiments that show the failure of near-threshold image distances to predict supra-threshold image distances are cited in [Chandler and Hemami 2003]. For the same reason, the simple root mean squared distance performed fairly well to image guided simplification, as compared to more sophisticated perceptual distances [Lindstrom and Turk 2000]. The Structural Similarity Index (SSIM) was recently proposed to avoid the supra-threshold problem for image perceptual distances [Wang et al. 2004]. SSIM is designed to extract structural information from the images, comparing separately the luminance, contrast, and structural components.

Several automatic distances, both based on images and geometry, are evaluated in [Watson et al. 2001]. In this study, the image-based automatic distances, namely the image mean-squared distance and the Bolin-Meyer metric [Bolin and Meyer 1998], use one view of the models to compute the distance between them. The geometric distances are based on 3D mesh data and are computed by the Metro tool [Cignoni et al. 1998]. The distances are evaluated against three experimental measures: forced choice preferences, ratings, and naming times. The study shows that human preferences and rankings correlate well to image mean-squared distance, Bolin-Meyer and Metro mean. The image-based distances proved to be better than geometric distances to predict human perception. The opportunity of using image-based distances for 3D models, as well as the importance of the illumination model, is discussed by [Rogowitz and Rushmeier 2001]. One conclusion is that people perceive still images differently than animations of the same models.

Several studies preferred human ratings as experimental measures [Rushmeier et al. 2000; Watson et al. 2001; Cheng et al. 2005]. However, analysis of rating data can easily fail. Too many rating levels have negative impact on the experiment [Luebke et al. 2002; Falmagne 1985], and the ordinal nature of ratings should be correspondingly analyzed [Krantz et al. 1971; Falmagne 1985; Rogowitz and Rushmeier 2001]. Often, experimental studies do not meet these requirements.

The main contribution of this paper is a new experiment inspired from the bisection method [Falmagne 1985] to model the supra-threshold perception of 3D models. Our approach avoids the problems of ratings experiments. We discuss two stochastic frameworks to evaluate the experimental data and investigate six automatic measures. We interpret the results and suggest further improvements.

2 Design of Experiment

One common element of the experiments on supra-threshold 3D perception is that they only consider the relations (such as ratings) between the original model (reference) and other models from its LOD hierarchy. Considering the perceptual metric space of the LOD set, such relations are not representative, as they do not sample the metric space uniformly [Suppes et al. 1989]. Instead, we evaluated relations among randomly sampled models from the LOD sequence. A similar approach is provided by the Maximum Likelihood Difference Scaling (MLDS) [Maloney and Yang 2003]. For

*e-mail: ioan.cleju@inf.uni-konstanz.de
†e-mail: dietmar.saupe@inf.uni-konstanz.de

the MLDS experiment, participants are asked to compare distances between two pairs of stimuli, sampling thus a quaternary relation on the studied space. MLDS was used to map a sequence of progressively compressed images onto a segment, preserving perceptual differences. A drawback of MLDS is the needed assumption of uni-dimensionality of the perceptual space, which we consider to be too restrictive in our case. Also, we believe the experiment to be too difficult for the users.

Our experiment used the bisection method to investigate the LOD perceptual space. For this technique, a participant is shown two stimuli and has to find a third stimulus whose effect is the average of the other two stimuli. In our study, we showed participants two models from the LOD sequence and asked them to find a third model at half-way perceptual distance between the two stimuli models. The users were explained that the half-way perceptual distance model should induce the maximum indecision on its similarity to either of the stimuli models. We concluded that, according to each participant, the perceptual distances between the two fixed models and the third model were equal. We presumed that the choices of the participants follow the normal distribution around the true perceptual half-way model.

We built a dense LOD sequence using the QSlim simplification algorithm [Garland and Heckbert 1997]. We assumed that the perceptual space containing the LOD sequence is a high dimensional Euclidean space. The assumption is not restrictive, since any metric can be embedded in a Euclidean space, given enough dimensions [Suppes et al. 1989]. The goal of the experiment is to compare the unknown metric of the perceptual space to the automatic metrics. Relations from the perceptual space were acquired by the user experiment. Ordinal analysis of the automatic measures was not effective because most of the measures sorted the set, revealing the same order as the LOD sequence.

We evaluated six automatic distances between 3D models, grouped in image-based and geometric distances. The image-based distances computation followed the framework of [Lindstrom and Turk 2000]: for each model, we rendered 8 images from uniformly distributed view points. For a comparison between two models, the distance is the average of the 8 corresponding image distances. Considering other studies on the perception of 3D models, we chose three image-based distances as candidates for supra-threshold perceptual measures of 3D models: structural similarity index (SSIM), mean of absolute differences of the pixels (IMn), and root mean squared distance (IRMS). As candidates for geometric supra-threshold perceptual distances we chose geometric root mean squared distance (MRMS) and mean geometric distance (MMn); we included the Hausdorff geometric distance (MHAUS) for its wide use in graphics community. All mentioned geometric distances were computed with the Metro tool.

2.1 Experimental Setup

For each reference model, we generated a LOD series consisting of 80 simplified models, ranging from 2% to 99% of the number of vertices of the original. The numbers of vertices of the simplified models were sampled equidistantly on a logarithmic scale. Each reference model consisted of about 2500 vertices. We obtained a discrete LOD set with small, approximatively uniform differences between two consecutive models.

The user interface is simple and easy to use. It contains three viewing panels, two for the stimuli models and one for the model to be selected. A slider component allows the participant to choose the model between the given stimuli, from the LOD sequence. The models are rendered in square panels of 400 pixels in length and

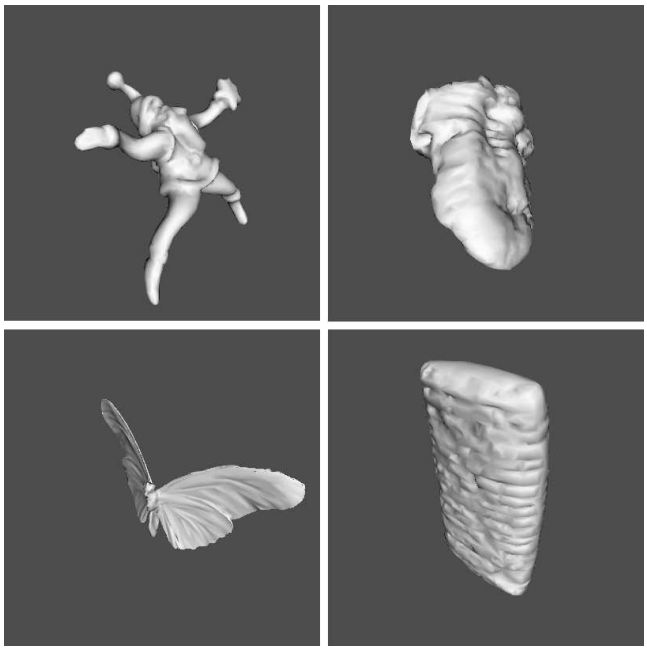

Figure 1: Models used for the experiment: Santa, Trilobite, Butterfly, and Cuneiform.

can be rotated and zoomed. Models are lit from front and Phong shaded, both for user interface and for image-based metrics. The panels are synchronized and allow the user to compare the three models under the same viewing conditions. Once the user decides about the selection, the next pair of stimuli is requested.

Nine volunteer participants took part in the experiment. We used 4 reference models with different geometric characteristics, namely Santa, Trilobite, Butterfly and Cuneiform, see Fig. 1. The Santa model was provided by Cyberware and the others by Arius3D. 15 random pairs of models were chosen for the stimuli for each reference model, giving a total of 60 trials for one participant. Each participant was given the instructions about the experiment and the user interface. 8 separate trials were used for training, 2 for each reference model. No time limits were imposed for decisions.

3 Experiment Analysis

The data of automatic measures consists of all-to-all distances between the models from each LOD set. For each user, experimental data contains a series of 60 triads, 15 for each reference model. Each triad contains indices of 3 models: the two fixed models and the user selection. We propose two probabilistic frameworks to compute the likelihood between each automatic metric and the experimental data: the a priori model and the a posteriori model.

The a priori model uses each automatic distance to predict the experimenter's choice. Let M_1 and M_2 be the stimuli and m_1,.., m_n the possible choices. For each choice m_i, we want to estimate the probability to be realized. Each triad (M_1,M_2,m_i) defines a plane in the Euclidean space given by an automatic metric $d()$. For each automatic metric $d()$, we define $D(d(),M_1,m_i,M_2)$ to be the signed distance from m_i to the perpendicular bisector of the segment M_1M_2, see Fig. 2. D can be easily computed from the known distances $d(M_1,m_i)$, $d(M_2,m_i)$ and $d(M_1,M_2)$. We model the distribution of D that predicts the users choices by a Gaussian with mean 0

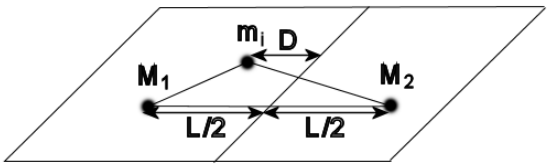

Figure 2: The two-dimensional subspace defined by the models M_1,m_i,M_2, embedded in the Euclidean space of an automatic metric.

and $d(M_1,M_2)/8$ standard deviation ($L/8$ in Fig. 2). We sample the probability density function (pdf) for each possible choice $m_1,...,m_n$ and normalize them to sum 1. The result is the probability of each possible choice to be realized. The likelihood that the choices of the users are a realization of the stochastic model is the product of the probabilities of the individual choices. Due to the reasonable scale, we show the corresponding logarithms (log-likelihoods), in Table 1.

For the a posteriori model, we keep the same definition of $D(d(),M_1,m_i,M_2)$. For each triad, each user choice $m_i = m$ gives a value for $D(d(),M_1,m,M_2)$. We estimate the distribution of $D(d(),M_1,:,M_2)$ from experimental data, under the assumption that it is Gaussian. The likelihood that the choice of the metric $d()$ is according to experimental data is the value of the pdf in 0. Similarly to the a priori likelihood, the a posteriori likelihood for the whole experiment is the product of the individual likelihoods. The individual likelihoods are not probabilities and can be larger than 1. Table 2 shows the mean values of log-likelihoods for the experiment.

The a posteriori model is clearly the better choice if the experimental data is sufficient. It incorporates the variance of users choices for each different trial in the experiment. The parameters of the a priori model are based entirely on automatic metrics data. It is more conservative and considers the same variance of user choices for all trials. We show the results for both scenarios.

Table 3 presents the ANOVA F-ratio values [Lindman 1991]. ANOVA tests the hypothesis that the means of the analyzed variables are the same for all the factors. We applied ANOVA to test if the individual log-likelihoods have the same mean for all models; we analyzed the D variable in the same way. Where possible, the log-likelihood and D were tested separately for all users, and averaged for each trial. For the geometric Hausdorff distance, D could not always be modelled by a Gaussian because the variance was 0 in some cases. This happened because for some stimuli (M_1,M_2), all participants choices $\{m_j,...,m_k\}$, were at the same distance to the stimuli: $d(M_1,m_j) = ... = d(M_1,m_k)$ and $d(M_2,m_j) = ... = d(M_2,m_k)$. Tables 2 and 3 do not provide values for these cases.

Metric	Santa	Trilobite	Butterfly	Cuneiform
SSIM	-3.23	-2.81	-2.92	-2.86
IMn	-2.99	-3.01	-2.91	-2.94
IRMS	-2.97	-3.02	-3.00	-2.96
MRMS	-4.06	-4.10	-3.90	-3.74
MMn	-3.95	-3.98	-3.79	-3.59
MHAUS	-4.05	-4.16	-3.97	-3.65

Table 1: A priori framework. Log-likelihood mean values for each metric and model.

Metric	Santa	Trilobite	Butterfly	Cuneiform
SSIM	-1.64	1.65	1.30	0.94
IMn	1.09	1.06	1.53	0.11
IRMS	1.39	0.87	1.63	0.95
MRMS	-10.93	-10.13	-7.07	-7.97
MMn	-8.32	-8.01	-4.93	-4.92
MHAUS	-13.86	-	-	-37.47

Table 2: A posteriori framework. Log-likelihood mean values for each metric and model.

Metric	**A priori**		**A posteriori**	**Deviation D**	
	Log-Lik	Log-Lik(avg)	Log-Lik	D	D(avg)
SSIM	10.30	2.30	7.82	89.93	26.92
IMn	0.60	0.11	1.02	17.61	4.97
IRMS	0.40	0.06	1.78	20.50	8.53
MRMS	2.79	0.49	0.34	3.18	0.44
MMn	2.93	0.54	0.50	1.75	0.27
MHAUS	5.60	0.93	-	10.3	1.32

Table 3: ANOVA F-ratio values for the factor model, for each metric. Avg indicates that the values are averaged for users.

3.1 Results

In all studied cases, the image-based metrics were clearly ranked better than geometric metrics. The differences between likelihoods were more evident for the a posteriori stochastic model, indicating that the assumption on fixed standard deviation for the a priori model was too conservative. The ranks of the image-based metrics depend on the model type and probabilistic model. According to both probabilistic frameworks, the order of image-based metrics was IRMS, IMn, and SSIM for the Santa model. For the model Trilobite, the order was SSIM, IMn and IRMS. The other two models were ranked differently by the two frameworks, but the absolute differences were not always significant. Metro Mean was better than Metro RMS and both of them were better than geometric Hausdorff distance. We confirmed that the average distances, such as geometric mean distance (MMn) and geometric root mean squared distance (MRMS) model better the perception than peak distances, such as geometric Hausdorff distances. ANOVA showed that the image-based metrics, especially SSIM, are more sensitive to model type than the geometric metrics.

The experimental data was not sufficient to rank the studied image-based distances. Considering also the ANOVA test, we believe that the performance of image-based distances is dependent on the geometrical particularities of the models and their behavior on simplification.

One important notice we made is that the geometric metrics are biased. Participants made choices consistently towards the better model of the pair, as compared to automatic methods. This suggests that the geometric-based distances (MMn and MRMS) are exaggerated for lower complexity models. One could improve the geometric distances in their perceptual behavior, weighting them based on the geometric complexity and the scale at which they will be rendered.

4 Future Work

We supported our experimental framework against ratings experiments with theoretic arguments. Our framework can bring more experimental data. It is much more difficult, if not impossible, to find the bias problem of the geometric metrics only with the help of

ratings experiments. We plan to find more experimental evidence to support the proposed framework against ratings. More case models should help us determine the performance of the different image-based metrics for classes of models.

The described experiment can be used to devise a technique to modify the geometric distances, in order to remove the bias. Data analysis shows that the geometric metrics have the advantage to be less sensible to the characteristics of the model; also, they do not depend on the illumination model. In this experiment we showed clear evidence that the studied image-based distances performed perceptually better than the geometric ones, but we also discussed ways to improve the geometric distances. The choice of an illumination model is a difficult issue for image-based distances and for designing user interfaces in perceptual studies. The task becomes more interesting as the metrics should be extended to textured models. However, considering the possible improvement of geometric distances that we discussed in the paper, we leave for future study the debate about the best type of perceptual metric for 3D models: image-based or geometric-based.

5 Acknowledgments

The work was supported by the DFG GK/1042 'Explorative Analysis and Visualization of Large Information Spaces' at the University of Konstanz, Germany. The Santa model was provided by Cyberware, and the Trilobite, the Butterfly, and the Cuneiform models were provided by Arius3D. SSIM code was taken from http://www.cns.nyu.edu/~lcv/ssim/, and we used the publicly available application Metro. We wish to thank the anonymous reviewers for their helpful comments and the volunteers to participate in the experiment.

References

BOLIN, M. R., AND MEYER, G. W. 1998. A perceptually based adaptive sampling algorithm. In *Proceedings of ACM SIGGRAPH 98*, ACM Press, 299–309.

CHANDLER, D. M., AND HEMAMI, S. S. 2003. Suprathreshold image compression based on contrast allocation and global precedence. In *Proceedings of SPIE Human Vision and Electronic Imaging VIII*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 5007, SPIE, 73–86.

CHENG, I., AND BOULANGER, P. 2005. A 3d perceptual metric using just-noticeable-difference. In *Eurographics Short Presentations*, 97–100.

CHENG, I., BASU, A., AND WANG, T. 2005. Balanced incomplete designs for 3d perceptual quality estimation. In *International Conference on Image Processing*, vol. 1, IEEE, 617–620.

CIGNONI, P., ROCCHINI, C., AND SCOPIGNO, R. 1998. Metro: Measuring error on simplified surfaces. *Computer Graphics Forum 17*, 2, 167–174.

FALMAGNE, J.-C. 1985. *Elements of Psychophysical Theory*. Claredon Press.

GARLAND, M., AND HECKBERT, P. S. 1997. Surface simplification using quadric error metrics. In *Proceedings of ACM SIGGRAPH 97*, ACM Press / Addison-Wesley Publishing Co., 209–216.

KRANTZ, D. H., LUCE, R. D., SUPPES, P., AND TVERSKY, A. 1971. *Foundations of Measurement*, vol. 1. Academic Press.

LINDMAN, H. R. 1991. *Analysis of Variance in Experimental Design*. Springer-Verlag.

LINDSTROM, P., AND TURK, G. 2000. Image-driven simplification. *ACM Transactions on Graphics 19*, 3, 204–241.

LINDSTROM, P. 2000. *Model Simplification using Image and Geometry-Based Metrics*. PhD thesis, Georgia Tech.

LUEBKE, D., AND HALLEN, B. 2001. Perceptually driven simplification for interactive rendering. In *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, Springer-Verlag, 223–234.

LUEBKE, D., REDDY, M., COHEN, J., VARSHNEY, A., WATSON, B., AND HUEBNER, R. 2002. *Level of Detail for 3D Graphics*. Morgan Kaufmann.

MALONEY, L. T., AND YANG, J. N. 2003. Maximum likelihood difference scaling. *Journal of Vision 3*, 8, 573–585.

REDDY, M. 1997. *Perceptually Modulated Level of Detail for Virtual Environments*. PhD thesis, University of Edinburgh.

REDDY, M. 2001. Perceptually optimized 3d graphics. *IEEE Computer Graphics and Applications 21*, 5, 68–75.

ROGOWITZ, B. E., AND RUSHMEIER, H. E. 2001. Are image quality metrics adequate to evaluate the quality of geometric objects? In *Proceedings of SPIE Human Vision and Electronic Imaging VI*, SPIE, B. E. Rogowitz and T. N. Pappas, Eds., vol. 4299, 340–348.

RUSHMEIER, H., ROGOWITZ, B., AND PIATKO, C. 2000. Perceptual issues in substituting texture for geometry. In *Proceedings of SPIE Human Vision and Electronic V*, vol. 3959, SPIE, 372–383.

SUPPES, P., KRANTZ, D. H., LUCE, R. D., AND TVERSKY, A. 1989. *Foundations of Measurement*, vol. 2. Academic Press.

WANG, Z., BOVIK, A. C., SHEIKH, H. R., AND SIMONCELLI, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4, 600–612.

WATSON, B., FRIEDMAN, A., AND MCGAFFEY, A. 2001. Measuring and predicting visual fidelity. In *Proceedings of ACM SIGGRAPH 2001*, ACM Press, ACM, 213–220.