Quality Assessment in Computer Graphics

Guillaume Lavoué and Rafał Mantiuk

Abstract In this chapter, we review the existing works regarding visual quality assessment in computer graphics. This broad area of research includes many sub-domains which make intensive use of quality assessment and/or artefact visibility evaluation: geometry processing, rendering, HDR imaging, tone mapping and stereo vision. For each of these sub-domains, we present the existing objective quality metrics, the subjective quality experiments as well as an evaluation and comparison of their performance. We broadly classify these existing works into image-based (i.e. evaluating artefacts introduced in 2D rendered images and videos) and model-based approaches (i.e. artefacts introduced on the 3D models themselves). Finally, the last part presents the emerging trends and main future directions.

1 Introduction

The realm of computer graphics is an intensive producer of visual content. Depending on the concerned sub-areas (e.g. geometric modeling, animation, rendering, simulation, high dynamic range (HDR) imaging, and so on) it generates and manipulates images, videos or 3D data. There is an obvious need to control and evaluate the quality of these graphical data regardless of the application. The term *quality* means here the *visual impact of the artefacts* introduced by the computer graphics techniques. For instance, in the context of rendering, one needs to evaluate the level of annoyance due to the noise introduced by an approximate illumination algorithm. As another example, for level of details creation, one needs to measure the visual impact of the simplification on the appearance of a 3D shape. Figure 1 illustrates

Guillaume Lavoué
University of Lyon, CNRS, Insa-Lyon, LIRIS UMR 5205, e-mail: glavoue@liris.cnrs.fr

Rafal Mantiuk
School of Computer Science, Bangor University, e-mail: mantiuk@gmail.com

these two examples of artefacts encountered in computer graphics. The paragraphs below introduce several useful terms that also point out the main differences between existing approaches for quality assessment in graphics.

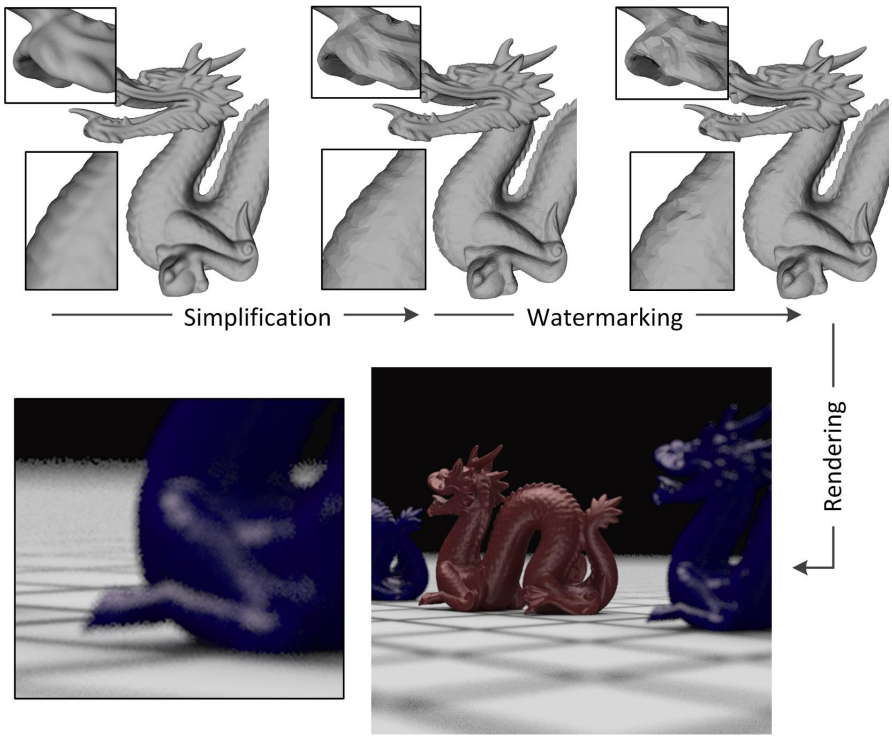

Fig. 1 Illustration of a typical computer graphics work-flow and its different sources of artefacts. *Top row, from left to right:* An original scanned 3D model (338K vertices); result after simplification (50K vertices) which introduces a rather uniform high frequency noise; result after watermarking [96] which creates some local bumps on the surface. *Bottom row:* Result after rendering (radiance caching) which introduces a non uniform structured noise.

Artefact visibility vs. global quality. For a given signal to evaluate (e.g. an image), the term *quality* often refers to a single score (mean-opinion-score) that aims at reflecting a kind of global level of annoyance caused by all artefacts and distortions in an image. Such global quality index is relevant for many computer graphics applications, e.g. to reduce/augment the sampling density in ray-tracing rendering. However, beside this global information, it is also important in many cases to obtain an information about the local *visibility* of the artefacts (i.e. predicting their spatial localization in the image). Such local information may allow, for instance, an automatic local corrections of the detected artefacts, like in [30].

Objective vs. sujective quality assessment. The quality evaluation of a given stimulus can be done directly by gathering the opinion of some observers by mean of a *subjective* experiment. However, this kind of study is obviously time-consuming, expensive and cannot be integrated into automatic processes. Hence researchers have focused on *objective* and automatic metrics that aim to predict this subjective visibility and/or quality. Both approaches are presented in this chapter.

Reference vs. no reference. Objective quality metrics can be classified according to the availability of the reference image (resp. video or 3D models): full-reference (FR), reduced reference (RR) and no-reference (NR). FR and RR metrics require at the quality evaluation stage that full or partial information on both images is present, the reference and the distorted one. NR metrics are much more challenging because they only have access to the distorted data; however, they are particularly relevant in computer graphics of which many techniques do not only *modify* but *create* visual content from abstract data. For instance, a rendering process generates a synthetic image from a 3D scene, hence to evaluate the rendering artefacts the metric will have access only to the test image since a perfect reference image without artefact is often unavailable.

Image artefacts vs. model artefacts. Computer graphics involves coarsely two main types of data: 3D data, i.e. surface and volume meshes issued from geometric modeling or scanning processes and 2D images and videos created/modified by graphical processes like rendering, tone mapping and so on. Usually, in a computer graphics work-flow (e.g. see figure 1), 3D data are first created (geometric modelling), processed (e.g. filtering, simplification) and then images/videos are generated from this 3D content (by rendering) and finally they can be post-processed (tone-mapped for instance). In such scenario, the visual defects at the very end of the processing chain may be due to artefacts introduced both on the 3D geometry (what we call model artefacts) and on the 2D image/video (what we called image artefacts). Since these two types of artefacts are introduced in very distinct processes and evaluated using very distinct metrics, each part of this chapter is divided according to this classification (except sections 2 and 3 respectively dedicated to each of them).

Black-box metrics vs. white-box metrics. There are two main approaches to modeling quality and fidelity: a black-box approach, which usually involves machine learning techniques; and a white-box approach, which attempts to model processes that are believed to exist in the human visual system. The visual difference predictors, such as VDP [20], are an example of a white-box approach, while the data-driven metrics for non-reference quality prediction [30] or color palette selection [66] are the examples of the black-box approach. Both approaches have their shortcomings. The black-box methods are good at fitting complex functions, but are prone to over-fitting. It is difficult to determine the right size of the training and testing data sets. Unless very large data sets are used, non-parametric models used in machine learning techniques cannot distinguish between major effects, which gov-

ern our perception of quality, and minor effects, which are unimportant. They are not suitable for finding a general patterns in the data and extracting a higher level understanding of the processes. Finally, the success of the machine learning methods depends on the choice of feature vectors, which need to be selected manually, relying in equal amounts on the expertise and a lucky guess.

White-box methods rely on the vast body of research devoted to modeling visual perception. They are less prone to over-fitting as they model only the effects that they are meant to predict. However, the choice of the right models is difficult. But even if the right set of models and right complexity is selected, combining and then calibrating them all together is a major challenge. Moreover, such white-box approaches are not very effective at accounting for higher level effects, such as aesthetics and naturalness, for which no models exist.

It is yet to be seen which approach will dominate and lead to the most successful quality metrics. It is also foreseeable that the metrics that combine both approaches will be able to benefit from their individual strengths and mitigate their weaknesses.

This chapter is organized as follows: sections 2 and 3 respectively present objective quality assessment regarding image artefacts and model artefacts. Then section 4 details the subjective quality experiments that have been conducted by the computer graphics community as well as quantitative evaluations of the objective metrics presented in sections 2 and 3. Finally section 5 is dedicated to the emerging trends and future research directions on the subject of quality assessment in graphics.

2 Image Quality Metrics in Graphics

2.1 Metrics for rendering based on visual models

Computer graphics rendering methods often rely on physical simulation of light propagation in a scene. Due to complex interaction of light with the environment and massive amount of light particles in a scene, these simulations require huge amount of computation. However, it has been long recognized that most applications of computer rendering methods require perceptually plausible solution rather than physically accurate results [72]. Knowing the limitations of the visual system, it should be possible to simplify the simulation and reduce the computational burden [67].

When rendering a scene, two important problems needs to be addressed: a) how to allocate samples (computation) over the image to improve perceptual quality; and b) when to stop collecting samples as further computation does not result in perceivable improvement. Both problems were considered in a series of papers on perceptually-based rendering, intended for both an accurate off-line techniques [11, 26, 12, 63, 73, 64, 65, 105, 30] as well as interactive rendering [54, 23]. Although the perceptual metrics used in these techniques operate in the image space, they are

different from the typical fidelity metrics, which compute the difference between reference and test images. Since the reference image is usually not available when rendering, these metrics aim at estimating error bounds based on approximation of the final image. Such an approximated image can be computed using fast GPU methods [105], by simulating only direct light (ray-casting) [73], approximating an image in the frequency domain [11, 12], using textures [95], intermediate rendering results [63], or consecutive animation frames [64]. Such an approximated images may not contain all the and illumination and shading details, especially those that are influenced by indirect lighting. However, the approximation is good enough to estimate the influence of both contrast and luminance masking in each part of the scene.

The visual metrics used in the rendering methods are predominantly based on visual difference predictors [20, 52], often extended to incorporate spatio-temporal contrast sensitivity function [34, 64, 65], opponent color processing and chromatic contrast sensitivity function (CSF) [62], and saliency models [31, 14]. Threshold versus elevation function [73, 23], photoreceptor non-linearity [63] or luminance-dependent CSF is used to model luminance masking, which accounts for the reduced sensitivity of the visual system at low luminance levels. Then, the image is decomposed into spatial-frequency and orientation selective bands using the Cortex transform [100, 63], wavelets [12], the DCT transform [95], or differences-of-Gaussians (DOGs) [26]. The spatial-sensitivity is incorporated either by pre-filtering the image with a CSF [63], or weighting each frequency band according to the CSF sensitivity value for its peak frequency [26, 73]. The multi-band decomposition is necessary to model contrast masking, which is realized either using a contrast transducer function [103, 26] or threshold elevation function [63, 73]. The visual difference predictors can be further weighted by a saliency map, which accounts for low-level attention [31, 73] and/or task-driven high-level attention [14].

In overall, the work on perceptual rendering influenced the way in which the perception is incorporated in graphics. Most methods in graphics rely on the near-threshold visual models and the notion of the just-noticeable-difference (JND). Such near-threshold models offer high accuracy and good rigor since the near-threshold models are well studied in the human vision research. But they also tend to result in over-conservative predictions and are not flexible enough to allow for visible but not disturbing distortions.

2.2 Open source metrics

The algorithms discussed for far incorporated visual metrics into rendering algorithms, making them difficult to test, compare or use as a fidelity metric on a pair of test and reference images. These metrics are also complex and hence challenging to reimplement with no source code publicly available. However, the graphics community have several alternative metrics to choose from if they wish to evaluate results without a need to reimplement visual models. *pdiff* [104] is a simple

perceptual difference metrics, which utilizes the CIE L*a*b* color space for differences in color, CSF and model of visual masking from Daly's VDP [20], and some speed improvements from [73]. The C source code is publicly available at http://pdiff.sourceforge.net/. A more complex visual model is offered by the series of HDR-VDP metrics [55, 56], which we discuss in more detail in Section 2.4. The older version of this metric (HDR-VDP-1.7.1) is available as a C/C++ code, while the latest version is provided as matlab sources (HDR-VDP-2.x). Both versions can be downloaded from http://hdrvdp.sf.net/.

2.3 Data-driven metrics for rendering

The majority of image metrics used in graphics rely on the models of the low-level visual perception. These metrics are often constructed by combining components from different visual models, such as saliency models, CSFs, threshold elevation functions and contrast transducers. While all these partial models well predict the individual effects, there is no guarantee that the combination of them will actually improve predictions. As shown in Section 4.4.1, complex metrics may actually perform worse in some tasks than a simple arithmetic difference. An alternative to such a white-box approach is the black-box approach, in which the metric is trained to predict differences based on a large data set. In this section we discuss two such metrics, one no-reference and one full-reference metric.

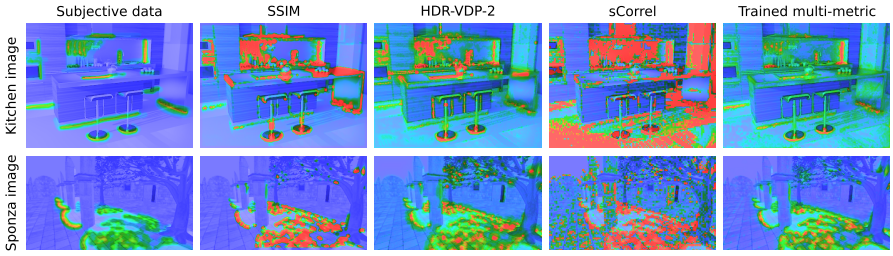

Fig. 2 Manually marked distortions in computer graphics rendering (left) and the predictions of image quality metrics: SSIM, HDR-VDP-2, sCorrel. Trained multi-metric uses the predictions of the existing metrics as a features for a decision forest classifier. It is trained to predict the subjective data.

Both metrics rely on the data collected in an experiment, in which observers were asked to label visible artefacts in computer graphics renderings, both when the reference image was shown and when it was hidden. The data set was the same as the one used to metric comparison, discussed in Section 4.4.1, though the non-reference metric was trained with only 10 images from that data set. Example of such manually marked images are shown in the left-most column in Figure 2. As compared to typical image quality databases, such as TID2008 [70], the maps of localized distortions provide much more data for the data-driven training. Instead of assigning

a single mean-opinion-score (MOS) to each image, the localized distortions maps provide up to a million of such numbers per image, as the labeling is associated with every image pixel. In practice a subsampled version of such a map is used because of limited accuracy of manual labeling. The limitation of such maps is that they do not provide the estimate of the perceived magnitude of distortion, which is given by MOS of DMOS. Instead, the maps contain the probability of detecting an artefact by an average observer.

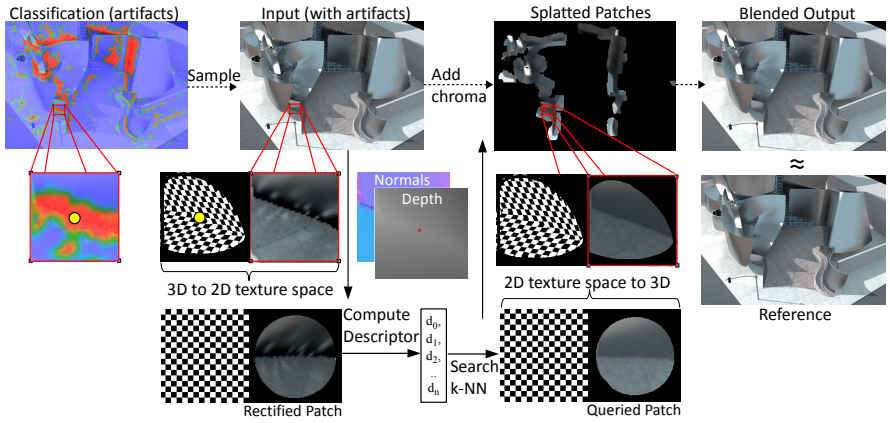

Fig. 3 Reduction of artefacts in rendered images by a metric-assisted inpainting [30]. Once the artefacts are detected in an image by a non-reference quality metric, the affected patches are replaced with similar non-distorted patches from the database. The operation is performed in an unfolded 2D texture space. The image courtesy of the authors.

Since a reference image is usually not available when rendering 3D scenes, Herzog et al. [30] proposed a no-reference image quality metric (NoRM) for three types of rendering distortions: VPL clamping, glossy VPL noise and shadow map aliasing. In contrast to other non-reference metrics, which can rely solely on a single color image, computer graphics method can provide additional information, such as a depth-buffer, or a diffuse material buffer. Such additional information was used alongside the color buffer to solve a rather challenging problem: predict visibility of artifacts given no reference image to compare with. The authors trained a support-vector-machine (SVM) based classifier on 10 images with manually labeled artefacts. The features used for training were an irradiance map with removed textures, screen-space ambient occlusion factor, unfolded textures described by the histogram of oriented gradients, a high-pass image with edges eliminated using the joint-bilateral filter and local statistics (mean, variance, skewness, kurtosis). Despite a rather small training set of 10 images, the metric was shown to provide comparable or better prediction performance than the state-of-the-art full reference metrics for the three types of targeted distortions. The authors describe also an application of this metric, in which detected artefacts are automatically corrected by inpainting.

The regions with detected artefacts are looked up a dictionary or artifact-free regions and replaced with a suitable substitute. The operation is illustrated in Figure 3.

The non-reference metrics are specialized in predicting only a certain kind of artefacts as they solve heavily under-constraint problem. Their predictive strength comes from learning the characteristic of a given artefacts and differentiating it from a regular image content. If a metric is to be used for a general purpose and with a wide variety of distortions, it needs to be supplied with both test and reference images.

Čadík et al. [90] explored a possibility of building a more reliable full-reference metric for rendering images using a data-driven approach. The motivation for this work was a previous study, showing mixed performance of existing metrics in this task (discussed in Section 4.4.1). They identified 32 image difference features, some described by a single number, some by up to 62 dimensions. Features ranged from a simple absolute difference to visual attention (measured with an eye-tracker) and included predictions of several major fidelity metrics (SSIM, HDR-VDP-2) and common computer vision descriptors (HOG, Harris corners, etc.). The metric was trained using 37 images with the manually labeled distortion maps. The best performance was achieved with ensembles of bagged decision trees (decision forest) used for classification. The classification was shown to perform significantly better than the best performing general purpose metric (sCorrel) as measured using the leave-one-out cross-validation procedure. Two examples of automatically generated distortion maps are shown in the right-most column of Figure 2 and compared with the predictions of other metrics.

Another example of a data-driven approach to metric design is the no-reference metric for evaluating the quality of motion deblurring, proposed by Liu et al. [51]. Motion deblurring algorithms aim at removing from photographs the motion blur due to camera shake. This is a blind deconvolution problem, in which the blur kernel is unknown. Since usually only blurry image is unavailable, it is essential to provide a mean to measure quality without the need for a sharp reference image. The data for training the metric was collected in a large scale crowd-sourcing experiment, in which over one thousand users ranked in a pairwise comparison experiments 40 scenes, each processed with 5 different deblurring algorithms. The metric was trained as a logistic regression explaining the relation between a number of features and the scaled subjective scores. The features included several no-reference measures of noise, sharpness, ringing and sharpness. In a dedicated validation experiment, the trained no-reference metric performed comparably or better than the state-of-the-art full reference metrics. The authors suggested several applications of the new metric, such as automatic selection of the deblurring algorithm which performs the best for a given image, or, on a local level, fusing high quality image by picking different image fragments from the result of each deblurring algorithm.

2.4 High dynamic range metrics for rendering

The majority of image quality metrics consider quality assessment for one particular medium, such as an LCD display or a print. However, the results of physically-accurate computer graphics methods are not tied to any concrete device. They produce images in which pixels contain linear radiometric values, as opposed to the gamma-corrected RGB values of a display device. Furthermore, the radiance values corresponding to real-world scenes can span a very large dynamic range, which exceeds the contrast range of a typical display device. Hence the problem arises of how to compare the quality of such images, which represent actual scenes, rather than their tone-mapped reproductions.

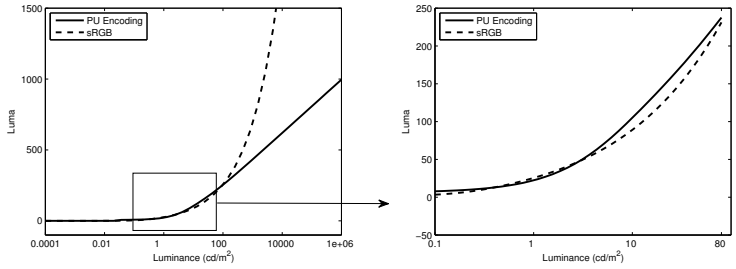

Fig. 4 Perceptually uniform (PU) encoding for evaluating quality of HDR images. The absolute luminance values are converted into luma values before they are used with standard image quality metrics, such as MSE, PSNR or SSIM. Note that the PU encoding is designed to give a good fit to the sRGB non-linearity within the range $0.1-80\ cd/m^2$ so that the results for low dynamic range images are consistent with those performed in the sRGB color space.

Aydin et al. [6] proposed a simple luminance encoding that makes it possible to use PSNR and SSIM [98] metrics with HDR images. The encoding transforms physical luminance values (represented in cd/m^2) into an approximately perceptually uniform representation (refer to Figure 4). The transformation is derived from luminance detection data using the threshold-integration method, similar to the one used for contrast transducer functions [103]. The transformation is further constrained so that the luminance values produced by a typical CRT display (in the range 0.1–$80\ cd/m^2$) are mapped to 0–255 range to mimic the sRGB non-linearity. This way, the quality predictions for typical low-dynamic range images are comparable to those calculated using pixel values. However, the metric can also operate in a much greater range of luminance.

The pixel encoding of Aydin et al. accounts for luminance masking, but it does not account for other luminance-dependent effects, such as inter-ocular light scatter or the frequency shift of the CSF peak with luminance. Those effects were modeled in the visual difference predictor for high dynamic range images (HDR-VDP) [55]. The HDR-VDP extends Daly's visual difference predictor (VDP) [20] to predict

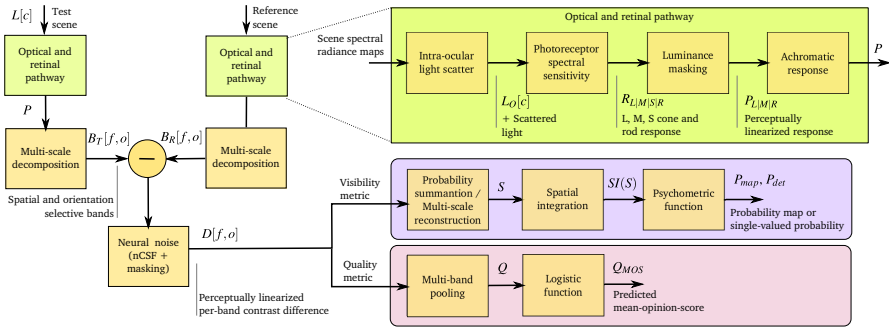

Fig. 5 The processing stages of the HDR-VDP-2 metric. Test and reference images undergo similar stages of visual modeling before they are compared at the level of individual spatial-and-orientation selective bands (B_T and B_R). The difference is used to predict both visibility (probability of detection) or quality (the perceived magnitude of distortion).

differences in high dynamic range images. In 2011 the metric was superseded with a completely redesigned metric HDR-VDP-2 [56], which is discussed below.

HDR-VDP-2 is the visibility (discrimination) and quality metric capable of detecting differences in achromatic images spanning a wide range of absolute luminance values [56]. Although the metric originates from the classical Visual Difference Predictor [20], and its extension — HDR-VDP [55], the visual models are very different from those used in those earlier metrics. The metric is also an effort to design a comprehensive model of the contrast visibility for a very wide range of illumination conditions.

As shown in Figure 5, the metric takes two HDR luminance or radiance maps as input and predicts the probability of detecting a difference between the pair of images (P_{map} and P_{det}) as well as the quality (Q and Q_{MOS}), which is defined as the perceived level of distortion.

One of the major factors limiting the contrast perception in high contrast (HDR) scenes is the scattering of the light in the optics of the eye and on the retina[59]. The HDR-VDP-2 models it as a frequency-space filter, which was fitted to an appropriate data set (*inter-ocular light scatter* block in Figure 5). The contrast perception deteriorates at lower luminance levels, where the vision is mediated mostly by night-vision photoreceptors — rods. This is especially manifested for small contrasts, which are close to the detection threshold. This effect is modeled as a hypothetical response of the photoreceptor (in steady state) to light (*luminance masking* block in Figure 5). Such response reduces the magnitude of image difference for low luminance according to the contrast detection measurements. The masking model (*neural noise* block in Figure 5) operates on the image decomposed into multiple orientation-and-frequency-selective bands to predict the threshold elevation due to contrast masking. Such masking is induced both by the contrast within the same band (intra-channel masking) and within neighboring bands (inter-channel masking). The same masking model incorporates also the effect of neural CSF, which is the contrast sensitivity function without the sensitivity reduction due to interocular

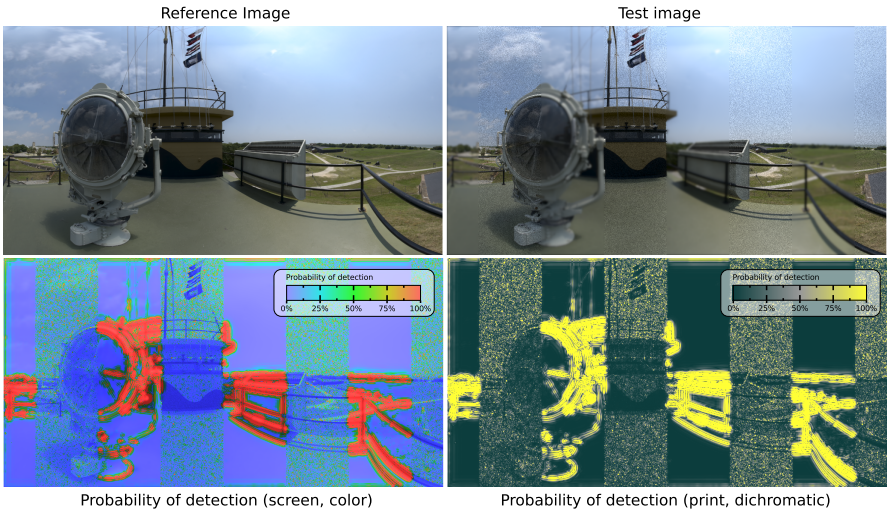

Fig. 6 Predicted visibility differences between the test and reference images. The test image contains interleaved vertical stripes of blur and white noise. The images are tone-mapped versions of an HDR input. The two color-coded maps on the right represent the probability that an average observer will notice a difference between the image pair. Both maps represent the same values, but use different color maps, optimized either for screen viewing or for gray-scale/color printing. The probability of detection drops with lower luminance (luminance sensitivity) and higher texture activity (contrast masking). Image courtesy of HDR-VFX, LLC 2008.

light scatter. Combining neural CSF with masking model is necessary to account for contrast constancy, which results in "flattening" of the CSF at the super-threshold contrast levels [27].

Figure 6 demonstrates the metric prediction for blur and noise. The model has been shown to predict numerous discrimination data sets, such as ModelFest [99], historical Blackwell's t.v.i. measurements [9], and newly measured CSF [35]. The source code of the metric is freely available for download from http://hdrvdp.sourceforge.net. It is also possible to run the metric using an on-line web service at http://driiqm.mpi-inf.mpg.de/.

2.5 Tone-mapping metrics

Tone mapping is the process of transforming an image represented in approximately physically accurate units, such as radiance and luminance, into pixel values that can be displayed on a screen of a limited dynamic range. Tone-mapping is a part of an image processing stack of any digital camera. A "raw" images captured by a digital sensor would produce unacceptable results if they were mapped directly to pixel values without any tone-mapping. But similar process is also necessary for all computer graphics methods that produce images represented in physical units. Therefore, the

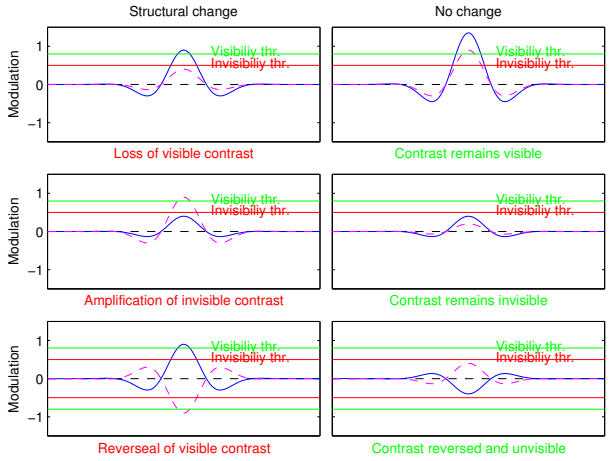

Fig. 7 The dynamic range independent metric [5] distinguished between the change of contrast that does and does not result in structural change. Blue continuous line shows a reference signal (from a band-pass pyramid) and magenta dashed line the test signal. When contrast remains visible or invisible after tone-mapping, no distortion is signalized (top and middle right). However, when the change of contrast alters the visibility of details, making visible details becoming invisible (top-left), it is signalized as a distortion.

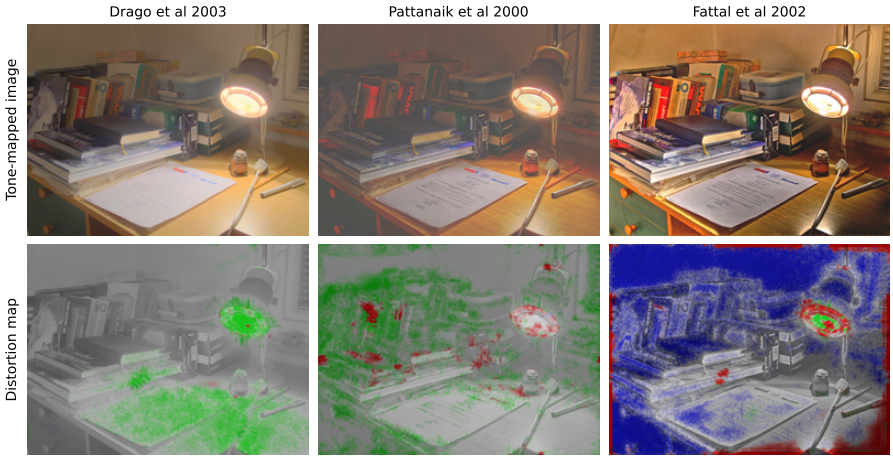

Fig. 8 Prediction of the dynamic range independent metric [5] (top) for tone-mapped images (bottom). The green color denotes the loss of visible contrast, the blue color the amplification of invisible contrast and the red color is contrast reversal (refer to Figure 7).

problem of tone-mapping and the quality assessment of tone-mapping results have been extensively studied in graphics.

Tone-mapping inherently produces images that are different from the original high dynamic range reference. In order to fit the resulting image within available color gamut and dynamic range of a display, tone-mapping often needs to compress

contrast and adjust brightness. Tone-mapped image may lose some quality as compared to the original seen on a high dynamic range display, yet the images look often very similar and the degradation of quality is poorly predicted by most quality metrics. Smith et al. [83] proposed the first metric intended for predicting loss of quality due to local and global contrast distortion introduced by tone-mapping. However, the metric was only used in the context of controlling counter-shading algorithm and was not validated against experimental data. Aydin et al. [5] proposed a metric for comparing HDR and tone-mapped images that is robust to contrast changes. The metric was later extended to video [7]. Both metrics are invariant to the change of contrast magnitude as long as that change does not distort contrast (inverse its polarity) or affect its visibility. The metric classifies distortions into three types: loss of visible contrast, amplification of invisible contrast and contrast reversal. All three cases are illustrated in Figure 7 on an example of a simple 2D Gabor patch. These three cases are believed to affect the quality of tone-mapped images. Figure 7 shows the metric predictions for three tone-mapped images. The main weakness of this metric is that produced distortion maps are suitable mostly for visual inspection and qualitative evaluation. The metric does not produce a single-valued quality estimate and its correlation with subjective quality assessment has not been verified.

Yeganeh and Wang [106] proposed a metric for tone mapping, which was designed to predict on overall quality of a tone-mapped image with respect to an HDR reference. The first component of the metric is the modification of the SSIM [98], which includes the contrast and structure components, but does not include the luminance component. The contrast component is further modified to detect only the cases in which invisible contrast becomes visible and visible contrast becomes invisible, in a similar spirit as in the dynamic range independent metric [5], described above. This is achieved by mapping local standard deviation values used in the contrast component into detection probabilities using a visual model, which consists of a psychometric function and a contrast sensitivity function (CSF). The second component of the metric describes "naturalness". The naturalness is captured by the measure of similarity between the histogram of a tone-mapped image and the distribution of histograms from the database of 3000 low-dynamic range images. The histogram is approximated by the Gaussian distribution. Then, its mean and standard deviation is compared against the database of histograms. When both values are likely to be found in the database, the image is considered natural and is assigned a higher quality. The metric was tested and cross-validated using three databases, including one from [92] and authors' own measurements. The Spearman rank-order correlation coefficient between the metric predictions and the subjective data was reported to be approximately 0.8. Such value is close to the performance of a random observer, which is estimated as the correlation between the mean and random observer's quality assessment.

Some visible distortions are desirable as long as they are not objectionable. An example of that is contrast enhancement through unsharp masking (high spatial frequencies) or countershading (low spatial frequencies) [37], commonly used in tone-mapping. In both cases, smooth gradients are introduced at both sides of an edge in order to enhance the contrast of that edge. This is demonstrated in Figure 9 where the

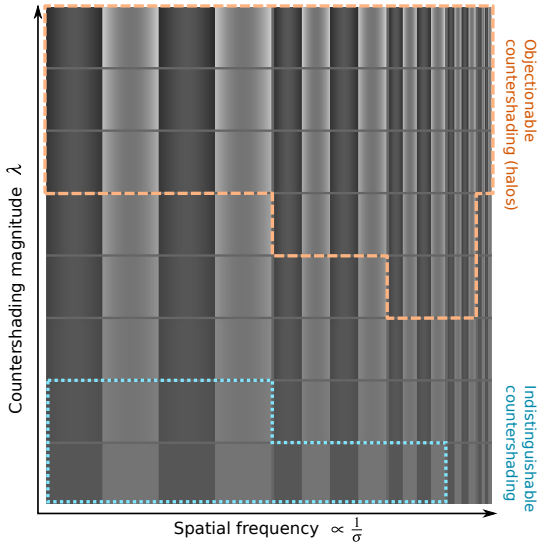

Fig. 9 Contrast enhancement by countershading. The figure shows the square-wave pattern with a reduced amplitude of the fundamental frequency, resulting in countershading profiles. The regions of indistinguishable (from a step edge) and objectionable countershading are marked with dotted and dashed lines of different color. The higher magnitude of countershading produces higher contrast edges. But if it is too high, the result appears objectionable. The marked regions are approximate and for illustration and actual regions will depend on the angular resolution of the figure.

base contrast shown in the bottom row is enhanced by adding contershading profiles. Note that the brightness of the central part of each patch remains the same across all rows. The region marked with the blue dashed line denotes the range of the Cornsweet illusion, where the gradient remains invisible while the edge is still enhanced. Above that line the Cornsweet illusion breaks and the gradients become visible. In practice, when countershading is added to tone-mapped images, it is actually desirable to introduce such visible gradients. Otherwise, the contrast enhancement is too small and does not improve image quality. But too strong gradient results in visible contrast reversal, also known as "halo" artifact, which is disturbing and objectionable. Trentacoste et al. [87] measured the threshold when countershading profiles become objectionable in complex images. They found that the permissible strength of the countershading depends on the width of the gradient profile, which in turn depends on the size of an image. They proposed a metric predicting the maximum strength of the enhancement and demonstrated its application to tone-mapping. The metric is an example of a problem where it is more important to predict when an artifact becomes objectionable rather than just visible.

2.6 Aesthetics and naturalness

Many quality assessment problems in graphics cannot be easily addressed by objective image and video metrics because they involve high level concepts, such as aesthetics or naturalness. For example, there is no computational algorithm that could tell whether an animation of a human character looks natural, or whether a scene composition looks pleasing to the eye. Yet, such tasks are often the goals of graphics methods. The common approach to such problems is to find a suitable set of numerical features that could correlate with subjective assessment, collect a large data set of subjective responses and then use machine learning techniques to train a predictor. Such methods proved to be effective for selecting the best viewpoint of a mesh[79], or selecting color palettes for graphic designs[66]. Yet, it is hard to expect that a suitable metric will be found for each individual problem. Therefore, graphics more often needs to rely on efficient subjective methods, which are discussed in Section 4.

3 Quality Metrics for 3D Models

The previous section focused on the quality evaluation of 2D images coming from computer graphics methods, mostly from rendering, HDR imaging or tone mapping. Hence most of the involved metrics aimed to detect specific image artefacts like aliasing, structured noise due to global illumination or halo artefacts from tone mapping. However, in computer graphics, visual artefacts do not come only from the final image creation process but they can occur on the 3D data themselves before the rendering. Indeed, 3D meshes are now subject to a wide range of processes which include transmission, compression, simplification, filtering, watermarking and so on. These processes inevitably introduce distortions which alter the geometry or texture of these 3D data and thus their final rendered appearance. Hence quality metrics have been introduced to detect these specific 3D artifacts, i.e. geometric quantization noise, smooth deformations due to watermarking, simplification artefacts and so on. A comprehensive review has been recently published about 3D mesh quality assessment [19]. Two kinds of approaches exist for this task: model-based and image-based approaches. Model-based approaches operate directly on the geometry and/or texture of the meshes being compared while image-based approaches considers rendered images of the 3D models (i.e. snapshots from different viewpoints) to evaluate their visual quality. Note that some image-based quality assessment algorithms consider only some specific viewpoints and thus are view-dependent.

3.1 Model-based metrics

In the fields of computer graphics, the first attempts to evaluate the visual fidelity of 3D objects were simple geometric distances, mainly used for driving mesh simplification [78]. A widely used metric is the Hausdorff distance, defined as follows:

$$H_a(M_1,M_2) = \max_{\mathbf{p}\in M_1} e(\mathbf{p},M_2) \tag{1}$$

with M_1 and M_2, the two 3D objects to compare and $e(\mathbf{p},M)$ the Euclidean distance from a point $\mathbf{p}$ in the 3D space to the surface M. This value is asymmetric; a symmetrical Hausdorff distance is defined as follows:

$$H(M_1,M_2) = \max\{H_a(M_1,M_2),H_a(M_2,M_1)\} \tag{2}$$

We can also define an asymmetric mean square error:

$$MSE_a(M_1,M_2) = \frac{1}{|M_1|}\int_{M_1} e(\mathbf{p},M_2)^2\,ds \tag{3}$$

The most widespread measurement is the Maximum Root Mean Square Error (MRMS):

$$MRMS(M_1,M_2) = \max\left\{\sqrt{MSE_a(M_1,M_2)},\sqrt{MSE_a(M_2,M_1)}\right\} \tag{4}$$

Cignoni et al. [16] provided the *Metro* software [1] with an implementation of Hausdorff and MRMS geometric distances between 3D models.

However these simple geometric measures are very poor predictor of the visual fidelity, like demonstrated in several studies [44, 89]. Hence, researchers have introduced perceptually-motivated metrics. These full reference metrics compare the distorted and original 3D models to compute a score which reflects the visual fidelity.

Karni and Gotsman [32], in order to evaluate properly their compression algorithm, consider the mean geometric distance between corresponding vertices and the mean distance of their geometric Laplacian values (which reflect a degree of smoothness of the surface) (this metric is abbreviated as $GL1$ in Table 1). Subsequently, Sorkine et al. [84] proposed a different version of this metric ($GL2$), which assumes slightly different values of the parameters involved. The performance of these metrics in terms of visual quality prediction remain low.

Several authors use the curvature information to derive perceptual quality metrics. Lavoué et al. [45] introduce the Mesh Structural Distortion Measure (MSDM) which follows the concept of structural similarity introduced for 2D image quality

[1] http://vcg.isti.cnr.it/activities/surfacegrevis/simplification/metro.html

assessment by Wang et al. [98] (SSIM index). The local *LMSDM* distance between two mesh local windows a and b is defined as follows:

$$LMSDM(a,b) = (\alpha L(a,b)^3 + \beta C(a,b)^3 + \gamma S(a,b)^3)^{\frac{1}{3}} \tag{5}$$

L, C and S represent respectively curvature, contrast and structure comparison functions:

$$\begin{aligned} L(a,b) &= \frac{\|\mu_a - \mu_b\|}{\max(\mu_a, \mu_b)} \\ C(a,b) &= \frac{\|\sigma_a - \sigma_b\|}{\max(\sigma_a, \sigma_b)} \\ S(a,b) &= \frac{\|\sigma_a \sigma_b - \sigma_{ab}\|}{\sigma_a \sigma_b} \end{aligned} \tag{6}$$

with μ_a, σ_a and σ_{ab} are respectively the mean, standard deviation and covariance of the curvature over the local windows a and b. A *local window* is defined as a connected set of vertices belonging to a sphere with a given radius. The global *MSDM* measure between two meshes is then defined by a Minkowski sum of the local distances associated with all local windows; it is a visual distortion index ranging from 0 (objects are identical) to 1 (theoretical limit when objects are completely different). A multi-resolution improved version, named *MSDM*2, has recently been proposed in [42]. It provides better performance and allows one to compare meshes with arbitrary connectivities. Torkhani et al. [86] introduced a similar metric called *TPDM* (Tensor-based Perceptual Distance Measure) which takes into account not only the mesh curvature amplitude, but also the principal curvature directions. Their motivation is that these directions represent structural features of the surface and thus should be visually important. These metrics owns the benefit of providing also a distortion map that predicts the perceived local artefacts visibility, like illustrated in figure 10.

Váša and Rus [89] consider the per-edge variations of oriented dihedral angles for visual quality assessment. The angle orientation allows to distinguish between convex and concave angles. Their metric (*DAME* for Dihedral Angle Mesh Error) is obtained by summing up the dihedral angle variations for all edges of the mesh being compared, as follows:

$$DAME = \frac{1}{n_e} \sum_{n_e} \|\alpha_i - \bar{\alpha}_i\| .m_i.w_i \tag{7}$$

with n_e the number of edges of the meshes being compared, α_i and $\bar{\alpha}_i$ the respective dihedral angles of the i^{th} edge of the original and distorted mesh. m_i is a weighting term relative to the masking effect (enhancing the distortion on smooth surfaces where they are the most visible). w_i is a weighting term relative to the surface visibility; indeed, a region almost always invisible should not contribute to the global distortion. This metric has the advantage of being very fast to compute but only

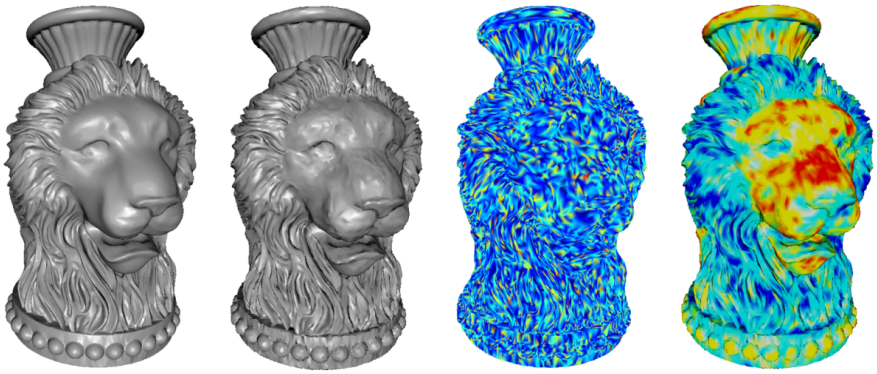

Fig. 10 *From left to right*: The Lion model; a distorted version after random noise addition; Hausdorff distortion map; MSDM2 distortion map. Warmer colors represent higher values.

works for comparing meshes of shared connectivity.

The metrics presented above consider local variations of attribute values at vertex or edge level, which are then pooled into a global score. In contrast, Corsini et al. [18] and Wang et al. [97] compute one global roughness value per 3D model and then derive a simple global roughness difference to derive a visual fidelity value between two 3D models. Corsini et al. [18] propose two ways of measuring the global model roughness. The first one is based on statistical considerations (at multiple scales) about the dihedral angles and the second one computes the variance of the geometric differences between a smoothed version of the model and its original version. These metrics are abbreviated as $3DWPM1$ and $3DWPM2$ in Table 1. Wang et al. [97] define the global roughness of a 3D model as a normalized surface integral of the local roughness, defined as the Laplacian of the discrete Gaussian curvature. The local roughness is modulated to take into account the masking effect. Their metric ($FMPD$ for Fast Mesh Perceptual Distance) provides good results and is fast to compute. Moreover a local distortion map can be obtained by differencing the local roughness values. Figure 11 illustrate some distorted versions of the *Horse* 3D model, with their corresponding $MRMS$, $MSDM2$ and $FMPD$ values.

Given the fact that all metrics above rely on different features e.g. curvature computations [45, 42, 86], dihedral angles [18, 89], Geometric Laplacian [32, 84], and Laplacian of Gaussian curvature [97]. Lavoué et al. [43] have hypothesized that a combination of these attributes could deliver better results that using them separately. They propose a quality metric based on an optimal linear combination of these attributes determined through machine learning. They obtained a very simple model which still provides good performance.

Some authors also proposed quality assessment metrics for textured 3D mesh [85, 68] dedicated to optimizing their compression and transmission. These metrics

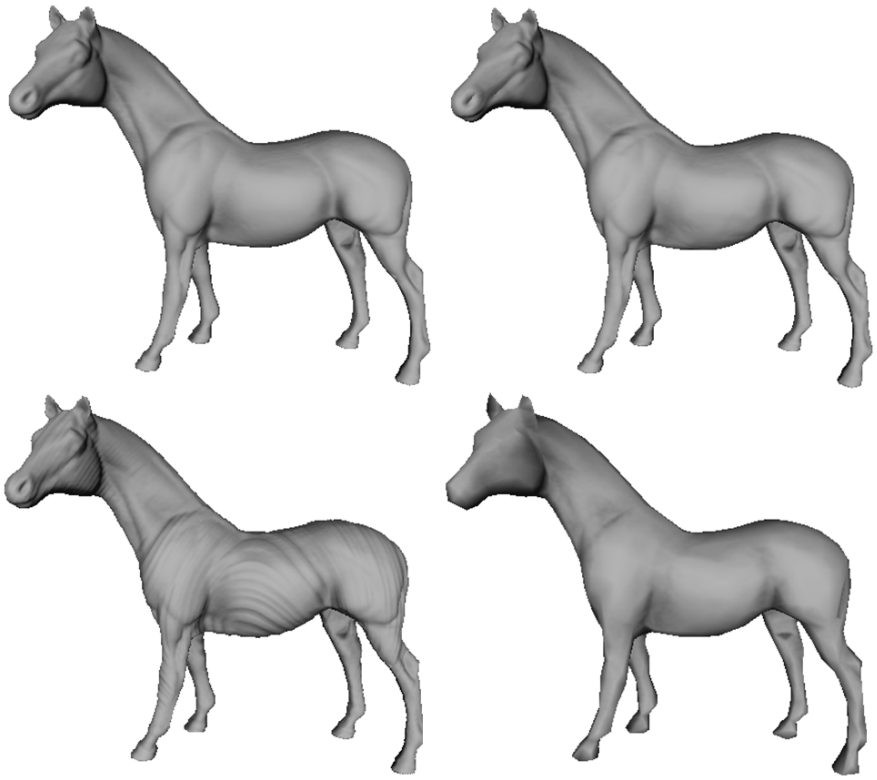

Fig. 11 Distorted versions of the *Horse* model, all associated with the same maximum root mean square error ($MRMS = 0.00105$). *From left to right, top to bottom*: Original model; result after watermarking from Wang et al. [96] (MSDM2=0.14, FMPD=0.01); result after watermarking from Cho et al. [15] (MSDM2=0.51, FMPD=0.40), result after simplification [49] from 113K vertices to 800 vertices (MSDM2=0.62, FMPD=1.00).

respectively rely on geometry and texture deviations [85] and on texture and mesh resolutions [68]. Their results underline the fact that the perceptual contribution of image texture is, in general, more important than the model's geometry, i.e. the reduction of the texture resolution is perceived more degraded than the reduction of model's polygons (geometry resolution).

For dynamic meshes, the most used metric is the KG error [33]. Given M_1 and M_2 the matrix representations ($3v \times f$ with v and f respectively the number of vertices and frames, 3 stands for the number of coordinates - x,y,z) of two dynamic meshes to compare, the KG error is defined as a normalized Frobenius norm of the matrix difference $\|M_1 - M_2\|$. Like the RMS for static meshes, this error metric does not correlate with the human vision. Váša and Skala have introduced a perceptual metric [88] for dynamic meshes, the STED error (Spatio-Temporal Edge Difference).

The metric works on edges as basic primitives, and computes the relative change in length for each edge of the mesh in each frame of the animation. This quality metric is able to capture both spatial and temporal artefacts and correlates well with the human vision.

Guthe et al. [28] introduce a perceptual metric based on spatio-temporal contrast sensitivity function dedicated to Bidirectional Texture Functions (BTFs), commonly used to represent the appearance of complex materials. This metric is used to measure the visual quality of the various compressed representations of BTF data.

Ramanarayanan et al. [72] proposed the concept of *visual equivalence* in order to create a metric that is more tolerant for non-disturbing artifacts. The authors proposed that two images are considered visually equivalent if object's shape and material are judged to be the same in both images and in a side-by-side comparison, an observer is unable to tell which image is closer to the reference. The authors proposed an experimental method and a metric (Visual Equivalence Predictor) based on the machine-learning techniques (SVM). The metric associates simple geometry and material descriptors with the samples measured in the experiments. Then, a trained classifier determines whether the distortions in illumination map leads to visually equivalent results. The metric demonstrated an interesting concept, yet it can be used only with a very limited range of illumination distortions. This work is dedicated to the evaluation of illumination map distortion effect, and not to the evaluation of the 3D model quality. However, it relies on geometry and material information and thus can be classified as a model-based metric.

3.2 Image-based metrics

Apart from these quality metrics operating on the 3D geometry (that we call model-based), a lot of researchers have used 2D image metrics to evaluate the visual quality of 3D graphical models. Indeed, as pointed out in [50], the main benefit of using image metrics to evaluate the visual quality of 3D objects is that the complex interactions between the various properties involved in the appearance (geometry, texture, normals) are naturally handled, avoiding the problem of how to combine and weight them. Many image-based quality evaluation works have been proposed in the context of simplification and level-of-detail (LoD) management for rendering. Among existing 2D metrics, authors have considered the Sarnoff Visual Discrimination Model (VDM) [52], the Visible Difference Predictor (VDP) from Daly [20] (both provide local distortion maps that predict local perceived differences), but also the SSIM (Structural SIMilarity) index, introduced by Wang and Bovik [98] and the classical mean or root mean squared pixel difference.

Lindstrom and Turk [50] evaluate the impact of simplification using a fast image quality metric (RMS error) computed on snapshots taken from 20 different camera

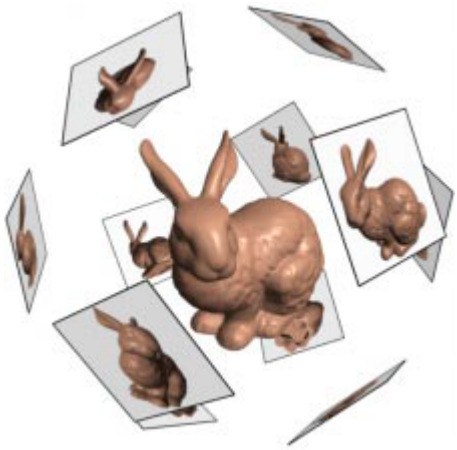

Fig. 12 Illustration of the image-based simplification approach from Lindstrom and Turk [50]. This algorithm considers the quality of 2D snapshots sampled around the 3D mesh as the main criterion for decimation. Image reprinted from [48].

positions regularly sampled on a bounding sphere. Their approach is illustrated in figure 12. In his PhD thesis [48], Lindstrom also proposed to replace the RMS by perceptual metrics including the Sarnoff VDM and surprisingly he found that the RMS error yields to better results. He also found that his image-based approach provides better results than geometry-driven approaches, however he considered a similar image-based evaluation. Qu and Meyer [71] consider the visual masking properties of 2D texture maps to drive simplification and remeshing of textured meshes, they evaluate the potential masking effect of the surface signals (mainly the texture) using the 2D Sarnoff VDM [52]. The masking map is obtained by comparing, using VDM, the original texture map with a Gaussian filtered version. The final remeshing can be view-independent or view-dependent depending on the visual effects considered. Zhu et al. [110] studied the relationship between the viewing distance and the perceptibility of model details using 2D metrics (VDP and SSIM) for the optimal design of discrete levels of detail (LOD) for the visualization of complex 3D building facades.

For animated characters, Larkin and O'Sullivan [40] ran an experiment to determine the influence of several types of artifacts (texture, silhouette and lighting) caused by simplification; they found that silhouette is the dominant artifact and then devised a quality metric based on silhouette changes suited to drive simplification. Their metric is as follow: they render local regions containing silhouette areas from different viewpoints and compare the resulting images with a 2D quality metric [104].

Several approaches do not rely directly on 2D metrics but rather on psychophysical models of visual perception (mostly the Contrast Sensitivity Function). One of the first study of this kind, was that of Reddy [74], which analyzed the frequency content in several pre-rendered images to determine the best LOD to use in a real-time rendering system. Luebke and Hallen [53] developed a perceptually-

based simplification algorithm based on a simplified version of the CSF. They map the change resulting from a local simplification operation to a worst-case contrast and a worst-case frequency and then determine whether this operation will be imperceptible. Their method was then extended by Williams et al. [102] to integrate texture and lighting effects. These latter approaches are view-dependent. Menzel and Guthe [61] propose a perceptual model of JND (Just Noticeable Difference) to drive their simplification algorithm; it integrates CSF and masking effect.The strength of their algorithm is to be able to perform almost all the calculation (i.e. contrast and frequency) directly on vertices instead of rendered images. However, it still uses the rendered views to evaluate the masking effect, thus it can be classified as an hybrid image-based/model-based method.

4 Subjective quality assessment in Graphics

Quality assessment metrics presented in sections 2 and 3 aim at *predicting* the visual quality and/or the local artefact visibility in graphics images and 3D models. Both these local and global perceived qualities can also be directly and quantitatively assessed by means of *subjective quality assessment experiments*. In such experiments, human observers give their opinion about the perceived quality or artefact visibility for a corpus of distorted images or 3D models.

Subjective experiments also provide a mean to test objective metrics. The non-parametric correlation, such as Spearman's or Kendall's rank-order correlation coefficients, computed between subjective scores and the objective scores provide an indicator of the performance of these metrics and a way to evaluate them quantitatively. We discuss some work in graphics on evaluation of objective metrics in Section 4.4.

For global quality assessment, many protocols exist and have been used for graphics data. Usually, absolute rating, double stimulus rating, ranking or pairwise comparisons are considered. Mantiuk et al. [57] compared the sensitivity and experiment duration for four experimental methods: single stimulus with a hidden reference, double stimulus, pairwise comparisons and similarity judgements. They found that the pairwise comparison method results in the lowest variation between observer's scores. Surprisingly, the method also required the shortest time to complete the experiments even for a large number of compared methods. This was believed to be due to the simplicity of the task, in which a better of two images was to be selected.

4.1 Scaling methods

Once experimental data is collected, it needs to be scaled into a mean a quality measure for a group of observers. Because different observers are likely to use different

scale when rating images, their results need to be unified. The easiest way to make their data comparable, is to apply a linear transform that makes the mean and the standard deviation equal for all observers. The result of such a transform is called z-score and is computed as

$$z_{i,j,k,r} = \frac{d_{i,j,k,r} - \bar{d}_i}{\sigma_i}, \tag{8}$$

where the mean score $\bar{d}_i$ and standard deviation σ_i are computed across all stimuli rated by an observer i. The indexes correspond to i-th observer, j-th condition (algorithm), k-th stimuli (image, video, etc.) and r-th repetition.

Pairwise comparison experiments require different scaling procedures, usually based on Thurstone Case IV or V assumptions [25]. These procedures attempt to convert the results of pairwise comparisons into a scale of Just-Noticeable-Differences (JND). When 75% of observers select one condition over another, the quality difference between them is assumed to be 1 JND. The scaling methods that tend to be the most robust are based on the maximum likelihood estimation [82, 3]. They maximize the probability that the scaled JND values explain the collected experimental data under the Thurstone Case V assumptions. The optimization procedure finds a quality value for each stimulus that maximizes the probability, which is modeled by the binomial distribution. Unlike standard scaling procedures, the probabilistic approach is robust to unanimous answers, which are common when a large number of conditions are compared. The detailed review of the scaling methods can be found in [25].

4.2 Specificity of graphics subjective experiments

4.2.1 Global vs. local

Artefacts coming from transmission or compression of natural images (i.e. blockiness, blurring, ringing) are mostly uniform. In contrast, artefacts from graphics processing or rendering are more often non uniform. Therefore, this domain needs visual metrics able to distinguish local artefacts visibility rather than global quality. Consequently, many experiments involving graphical content involve locally marking noticeable and objectionable distortions [91] rather than judging an overall quality. This marking task is more complicated than a quality rating, thus it involves the creation of innovative protocols.

4.2.2 Large number of parameters

A subjective experiment usually involves an number of important parameters; for instance, for evaluating the quality of images or videos, one has to decide the corpus

of data, the nature and amplitude of the distortions as well as the rating protocol (i.e. single or multiple stimulus, continuous or category rating, etc). However, the design of a subjective study involving 3D graphical content requires many additional parameters (as raised in [13]):

- Lighting. As raised in the experiment of Rogowitz and Rushmeier [75], the position and type of light source(s) have a strong influence on the perception of the artefacts.
- Materials and Shading. Complex materials and shaders may enhance the artefacts visibility, or on the contrary, act as a masker (in particular some texture patterns [26]).
- Background. The background may affect the perceived quality of the 3D model, in particular it influences the visibility of the silhouette, which strongly influences the perception.
- Animation & interaction. There exist different ways to display the 3D models to the observers, from the most simple (e.g. as a static image from one given viewpoint, as in [101]) to the most complex (e.g. by allowing free rotation, zoom, translation, as in [18]). Of course it is important for the observer to have access to different viewpoints of the objects, however the problem of allowing free interaction is the cognitive overload that may alter the results. A good compromise may be the use of animations, as in [68], however the velocity strongly influences the Contrast Sensitivity Function [34], hence animations have to be reasonably slow.

4.2.3 Specifics of tone-mapping evaluation

In this section we discuss the importance of selecting the right reference and an evaluation method for subjective evaluation of tone-mapping operators. This section serves as an example of the considerations that are relevant when considering quality assessment in graphics applications. Similar text has been published before in [24].

Figure 13 illustrates a general tone mapping scenario and a number of possible evaluation methods. To create an HDR image, the physical light intensities (luminance and radiance) in a scene are captured with a camera or rendered using computer graphics methods. In the general case, "RAW" camera formats can be considered as HDR formats, as they do not alter captured light information given a linear response of a CCD/CMOS sensor. In the case of professional content production, the creator (director, artist) seldom wants to show what has been captured in a physical scene. The camera-captured content is edited, color-graded and enhanced. This can be done manually by a color artist or automatically by color processing software. It is important to distinguish this step from actual tone-mapping, which, in our view, is meant to do "the least damage" to the appearance of carefully edited content. In some applications, such as simulators or realistic visualization, where faithful reproduction is crucial, the enhancement step is omitted.

Tone-mapping can be targeted for a range of displays, which may differ substantially in their contrast and brightness levels. Even HDR displays require tone-

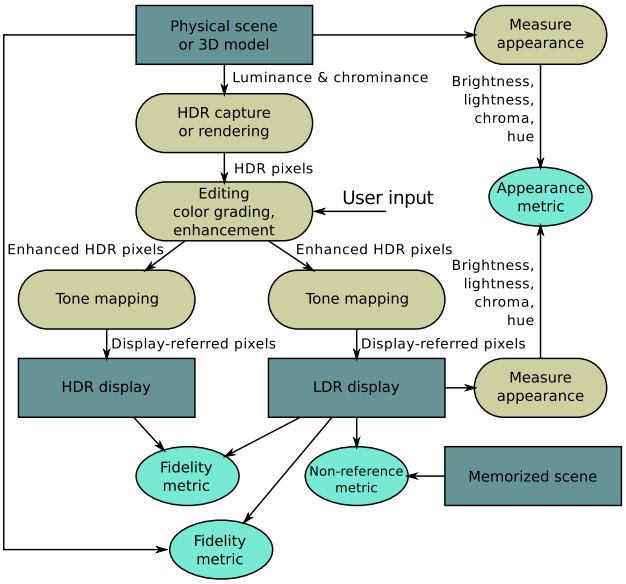

Fig. 13 Tone-mapping process and different methods of performing tone-mapping evaluation. Note that content editing has been distinguished from tone-mapping. The evaluation methods (subjective metrics) are shown as ovals.

mapping as they are incapable of reproducing the luminance levels found in the real world. An HDR display, however, can be considered as the best possible reproduction available, or a "reference" display. Given such a tone-mapping pipeline, we can distinguish the following evaluation methods:

Fidelity with reality method, where a tone-mapped image is compared with a physical scene. Such a study is challenging to execute, in particular for video because it involves displaying both a tone-mapped image and the corresponding physical scene in the same experimental setup. Furthermore, the task is very difficult for observers as displayed scenes differ from real scenes not only in the dynamic range, but they also lack stereo depth, focal cues, and have restricted field of view and color gamut. These factors usually cannot be controlled or eliminated. Moreover, this task does not capture the actual intent when the content needs enhancement. Despite the above issues, the method directly tests one of the main objectives of tone-mapping and was used in a number of studies [107, 4, 108, 92, 93].

Fidelity with HDR reproduction methods, where content is matched against a reference shown on an HDR display. Although HDR displays offer a potentially large dynamic range, some form of tone-mapping, such as absolute luminance adjustment and clipping, is still required to reproduce the original content. This introduces imperfections in the displayed reference content. For example, an HDR display will not evoke the same sensation of glare in the eye as the actual scene. However, the approach has the advantage that the experiments can be run in a well-controlled environment and, given the reference, the task is easier. Because of the

limited availability of HDR displays, only a few studies employed this method: [47, 38].

Non-reference methods, where observers are asked to evaluate operators without being shown any reference. In many applications there is no need for fidelity with "perfect" or "reference" reproduction. For example, the consumer photography is focused on making images look possibly good on a device or print alone as most consumers will rarely judge the images while comparing with real scenes. Although the method is simple and targets many applications, it carries the risk of running a "beauty contest" [60], where the criteria of evaluation are very subjective. In the non-reference scenario, it is commonly assumed that tone-mapping is also responsible for performing color editing and enhancement. But, since people differ a lot in their preference for enhancement [108], such studies lead to very inconsistent results. The best results are achieved if the algorithm is tweaked independently for each scene, or essentially if a color artist is involved. In this way we are not testing an automatic algorithm though, but a color editing tool and the skills of the artist. However, if these issues are well controlled, the method provides a convenient way to test TMO performance against user expectations and, therefore, it was employed in most of the studies on tone-mapping: [39, 21, 4, 108, 1, 92, 69].

Appearance match methods compare color appearance in both the original scene and its reproduction [60]. For example, the brightness of square patches can be measured in a physical scene and on a display using the magnitude estimation methods. Then, the best tone-mapping is the one that provides the best match between the measured perceptual attributes. Even though this seems to be a very precise method, it poses a number of problems. Firstly, measuring appearance for complex scenes is challenging. While measuring brightness for uniform patches is a tractable task, there is no easy method to measure the appearance of gloss, gradients, textures and complex materials. Secondly, the match of sparsely measured perceptual attributes does not need to guarantee the overall match of image appearance.

None of the discussed evaluation methods is free of problems. The choice of a method depends on the application that is relevant to the study. The diversity of the methods shows the challenge of subjective quality assessment in tone-mapping, and is one of the factors that contribute to volatility of the results.

4.2.4 Volatility of the results

It is not uncommon to find quality studies in graphics, which arrive with contradicting or inconclusive results. For example, two studies [8, 58] compared inverse tone-mapping operators. Both studies asked to rate or rank the fidelity of the processed image with the reference shown on an HDR display. The first study [8] demonstrated that the performance of complex operators is superior to that of a simple linear scaling. The second study [58] arrived with the opposite conclusion, that the linear contrast scaling performs comparably or better than the complex operators. Both studies compared the same operators, but images, parameter settings for each algorithm, evaluation methods and experimental conditions were different. This two

conflicting results show the volatility of many subjective experiments performed on images. The statistical testing employed in these studies can ensure that the results are likely to be the same if the experiment is repeated for a different group of observers, but with exactly the same images and in exactly the same conditions. The statistical testing, however, cannot generalize the results to the entire population of possible images, parameters, experimental conditions and evaluation procedures.

4.3 Subjective quality experiments

This subsection presents the subjective tests conducted by the scientific community related to quality assessment of graphics data. The first and second parts detail respectively experiments related to image and 3D model artefact evaluation.

4.3.1 Image and Video Quality Assessment

Evaluating computer graphics methods is inherently difficult, as the results can often be only evaluated visually. This poses a challenge for the authors of new algorithms, who are expected to compare their results with the state-of-the-art. For that reason, many recent papers in graphics include a short section with experimental validation. Such a trend shows that subjective quality assessment becomes a standard practice and a part of the research methodology in graphics. The need to validate methods also motivates comparative studies, in which several state-of-the-art algorithms are evaluated in a subjective experiment. Studies like this have been performed for image aspect ratio retargeting [76], image deghosting [29] or inverse tone-mapping [8, 58]. However, probably the most attention has attracted the problem of tone-mapping, which is discussed below.

Currently (as of 2014) Google Scholar search reports over 7,000 papers with the phrase “tone mapping” in the title. Given this enormous choice of different algorithms, which accomplish a very similar task, one would wish to know which algorithm performs the best in a general case. In Section 2.5 we discussed a few objective metrics for tone-mapping. However, because their accuracy still needs to be validated, they are not commonly recognized method for comparing tone-mapping operators. Instead, the operators have been compared in a large number of subjective studies evaluating both tone mapping for static images [22, 39, 47, 21, 107, 4, 108, 1, 2, 92, 38, 93, 36] and tone mapping for video [69, 24, 10]. None of these studies provided a definite ranking of the operators since such a ranking strongly depends on the scene content and the parameters passed to a tone-mapping operator. Interestingly, many complex tone-mapping methods seem to perform comparable or worse than even a simple method, provided that it is fine-tuned manually [1, 92, 38]. This shows the importance of per-image parameter tuning. Furthermore, the objective (intent) of tone mapping can be very different between operators. Some operators simulate the performance of the visual system with all its limitation; other

operators minimize color differences between the HDR image and its reproduction; and some produce the most pleasing images [24, 60]. Therefore, a single ranking and evaluation criteria do not seem to be appropriate for evaluation of all types of tone mapping. The studies have identified the factors that affect overall quality of the results, such as naturalness and detail [22], overall contrast and brightness reproduction [107, 108], color reproduction and visible artefacts [92]. In case of video tone mapping, the overall quality is also affected by flickering, ghosting, noise and consistency of colors across a video sequence [24, 10]. Evaluating all these attributes provides the most insight into the performance of the operators but it also requires the most effort and expertise and, therefore, is often performed by expert observers [24]. In overall, the subjective studies have not identified a single operator that would perform well a general case. But they helped to identify common problems in tone-mapping, which will help in guiding further research on this topic.

4.3.2 3D Model Quality Assessment

Several authors have made subjective tests involving 3D static or dynamic models [77, 101, 75, 68, 17, 45, 18, 80, 81, 41, 88, 89]. Their experiments, detailed below, had different purposes and used different methodologies. Bulbul et al. [13] recently provided a good overview and comparison of their environments, methodologies and materials.

Subjective tests from Watson et al. [101] and Rogowitz and Rushmeier [75] focus on a mesh simplification scenario; their test databases were created by applying different simplification algorithms at different ratios on several 3D models. They considered a double stimulus rating scenario, i.e. observers had to rate the fidelity of simplified models regarding the original ones. The purposes of their experiments were respectively to compare image-based metrics and geometric ones to predict the perceived degradation of simplified 3D models [101] and to study if 2D images of a 3D model are really suited to evaluate its quality [75].

Rushmeier et al. [77] and Pan et al. [68] also considered a simplification scenario; however, their 3D models were textured. These experiments provided useful insights on how resolution of texture and resolution of mesh influence the visual appearance of the object. Pan et al. [68] also provided a perceptual metric predicting this visual quality and evaluated it quantitatively by studying the correlation with subjective MOS from their experiment.

Lavoué [41] conducted an experiment involving 3D objects specifically chosen because they contain significantly smooth and rough areas. The author applied noise addition with different strengths either on smooth or rough areas. The specific objective of this study was to evaluate the *visual masking* effect. It turns out that the noise is indeed far less visible on rough regions. Hence, the metrics should follow this perceptual mechanism. The data resulting from this experiment (Masking Database

in table 1) are publicly available [2].

To the best of our knowledge, the only experiment involving dynamic meshes was the one performed by Váša and Skala [88] in their work proposing the STED metric. They considered 5 dynamic meshes (chicken, dance, cloth, mocap and jump) and applied different kinds of both spatial and temporal distortion of varying types: random noise, smooth sinusoidal dislocation of vertices, temporal shaking and results of various compression algorithms. All the versions (including the original) were displayed at the same time to the observers, and they were asked to rate them using a continuous scale from 0 to 10.

In all the studies presented above, the observers are asked to rate the fidelity of a distorted model regarding a reference one, displayed at the same time (usually a double stimulus scenario). However some experiments consider a *single stimulus absolute rating* scenario. Corsini et al. [18] proposed two subjective experiments focusing on a watermarking scenario; the material was composed of 3D models processed by different watermarking algorithms introducing different kinds of artifacts. On the contrary to the studies presented above, they consider an absolute rating with hidden reference (i.e. the reference is displayed among the other stimuli). The authors then used the mean opinion scores to evaluate the effectiveness of several geometric metrics and proposed a new perceptual one (see section 3.1) to assess the quality of watermarked 3D models. Lavoué et al. [45] follow the same protocol for their study; their material is composed of 88 models generated from 4 reference objects (Armadillo, Dyno, Venus and RockerArm). Two types of distortion (noise addition and smoothing) are applied with different strengths and non uniformly on the object surface. The resulting MOS were originally used to evaluate the performance of the MSDM perceptual metric (see section 3.1). The corresponding database (General-Purpose Database in table 1) and MOS data are publicly available [2].

Rating experiment have the benefit of directly providing a mean opinion score for each object from the corpus, however the task of assigning a quality score to each stimulus is difficult for the observers and may lead to inaccurate results. That is why many experiments now rely on the simpler task of *Paired Comparison* where observers just have to provide a preference between a pair of stimuli (usually as a binary forced choice). Silva et al. [80] proposed an experiment involving both rating and preference tasks. Their corpus contains 30 models generated from 5 reference objects. The reference models have been simplified using three different methods and two levels. For the rating task, observers were asked to provide a score from 1 (very bad) to 5 (very good). Along with this rating, in another phase of the test, the observers were asked about their preference among several simplified models presented together. Figure 14 illustrates the evaluation interface for the rating task, the stimulus to rate is presented with its reference stimulus. The data resulting from

[2] http://liris.cnrs.fr/guillaume.lavoue/data/datasets.html

these subjective experiments are publicly available [3] (Simplification Database in table 1). The same authors did another subjective experiment using a larger corpus of models [81] where they only collected preferences.

Váša and Rus [89] conducted a subjective study focusing on evaluating compression artefacts. Their corpus contains 65 models from 5 references. The applied distortions are uniform and Gaussian noise, sine signal, geometric quantization, affine transform, smoothing and results from three compression algorithms. The observer's task is a binary forced choice, in the presence of the reference; i.e. triplets of meshes were presented, with one mesh being designated as original, and two randomly chosen distorted versions. A scalar quality value for each object from the corpus is then derived from the user choices. The data (Compression Database in table 1) are publicly available [4].

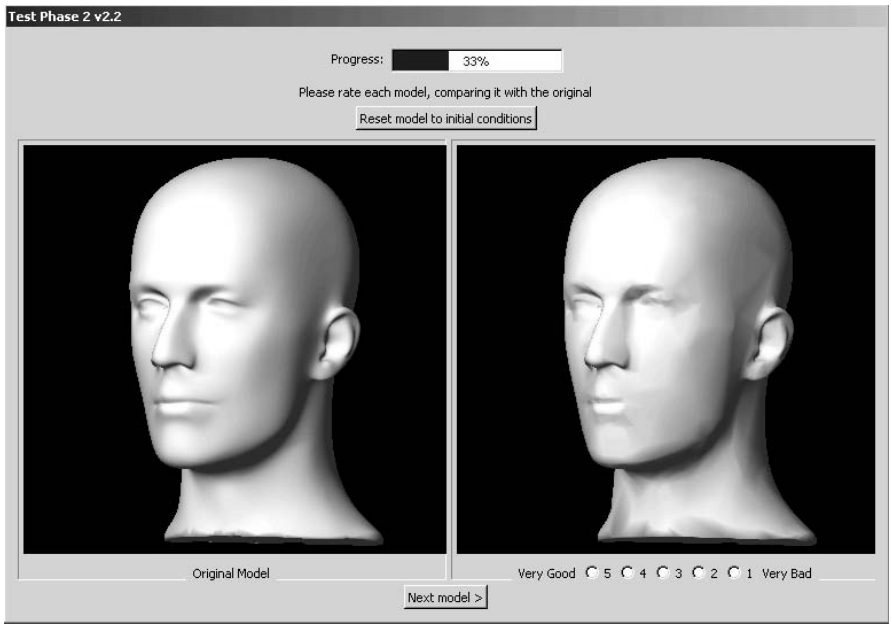

Fig. 14 Evaluation interface for the subjective test of Silva et al. [81]. The observers were asked to compare the target stimulus (right) with the referential stimuli (left) and assign it a category rating from 1 (very bad) to 5 (very good). Reprinted from [81].

[3] http://www.ieeta.pt/ sss/repository/

[4] http://compression.kiv.zcu.cz/

4.4 Performance of quality metrics

4.4.1 Image Quality Assessment for Rendering

VDP-like metrics are, which are dominant in graphics, often considered to be too sensitive to small, barely noticeable, and often negligible differences. For example, many computer graphics methods result in a bias, which makes the part of a rendered scene brighter or darker than the physically accurate reference. Since such a brightness change is local, smooth and spatially consistent, most observers are unlikely to notice it unless they scrupulously compare the image with a reference. Yet, such a difference will be signalized as significant by most VDP-like metrics, which will correctly predict that the difference is in fact visible when scrutinized. As a result, the distortion maps produced by objective metrics often do not correspond well with subjective judgment about visible artefacts.

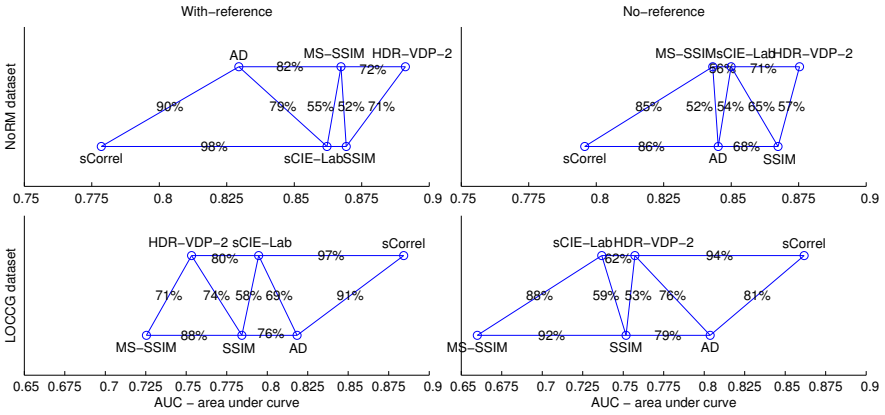

Fig. 15 The performance of quality metrics according to the area-under-curve (AUC) (the higher the AUC, the better the classification into distorted and undistorted regions). The top row shows the results for the NoRM data set [30] and bottom row the LOCCG data [91]. The columns correspond to the experiments in which the reference non-distorted image was shown (left column) or hidden (right column). The percentages indicate how frequently the metric on the right results in higher AUC when the image set is randomized using a bootstrapping procedure. The metrics: AD — absolute difference (equivalent to PSNR); SSIM - Structural Similarity Index; MS-SSIM — multi-scale SSIM; HDR-VDP-2 — refer to Section 2.4; sCIE-Lab — spatial CIELab; sCorrel — per-block Spearman's nonparametric correlation.

Cadík et al. [91] investigated this problem by comparing the performance of the state-of-the-art fidelity metrics in predicting rendering artefacts. The selected metrics were based on perceptual models (HDR-VDP-2), texture statistics (SSIM, MS-SSIM), color differences (sCIE-Lab) and simple arithmetic difference (MSE). The performance was compared against experimental data, which was collected by ask-

ing observers to label noticeable artifacts in images. Two examples of such manually labelled distortions maps are shown in Figure 2.

The same group of observers completed the experiment for two different tasks. The first task involved marking artifacts without revealing the reference (artifact free) image. It relied on the observers being able to spot objectionable distortions. In the second task the reference image was shown next to the distorted and the observers were asked to find all visible differences. The results for both tasks were mostly consistent across observers resulting in similar distortion maps for each individual.

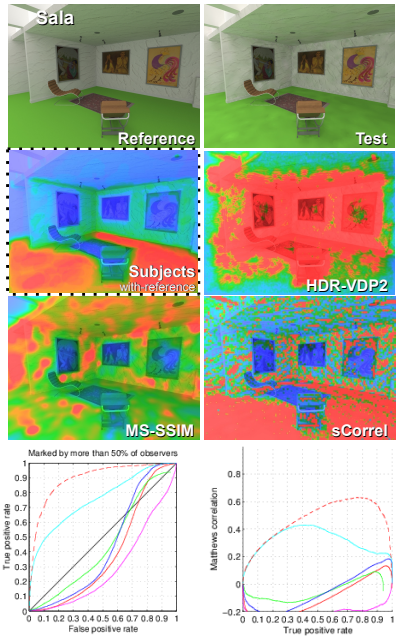

Fig. 16 Scene *sala* (top), distortion maps for selected metrics (2^{nd} and 3^{rd} rows), ROC and correlation plots (bottom). Most metrics are sensitive to brightness changes, which often remain unnoticed by observers. *sCorrel* (block-wise Spearson correlation) is the only metric robust to these artifacts. Refer to the legend in Figure 15 to check which lines correspond to which metrics in the plots.

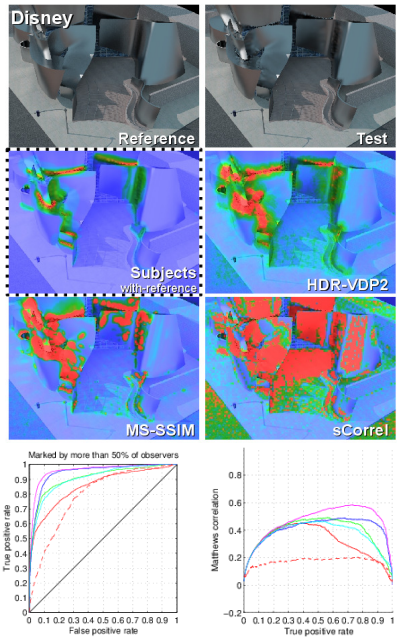

Fig. 17 Scene *disney*: simple metrics, such as sCorrel and AD, fail to distinguish between visible and invisible amount of noise resulting in worse performance.

When subjective distortion maps were compared against the metric predictions, they revealed weaknesses of both simple (PSNR, sCIE-Lab [109]) and advanced (SSIM, MS-SSIM [98], HDR-VDP-2) quality metrics. The results for the two separate data sets (NORM [30] and LOCCG[91]) and two experimental conditions (with-reference and no-reference) are shown in Figure 15. The results show that the

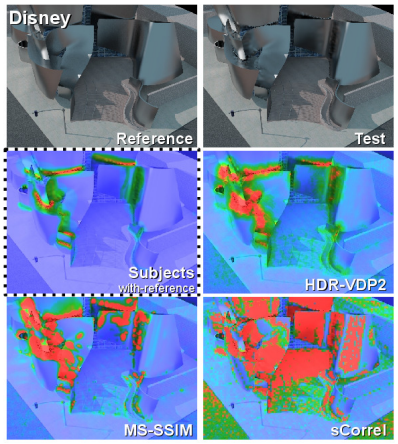

Fig. 18 *Dragons* scene contains artifacts on the dragon figures but not in the black background. Multi-scale IQMs, such as MS-SSIM and HDR-VDP-2, mark much larger regions due to the differences detected at lower spatial frequencies. Pixel-based AD (absolute differences) can better localize distortions in this case.

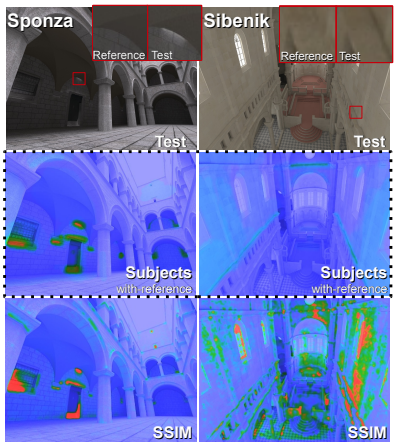

Fig. 19 Photon leaking and VPL clamping artifacts in scenes *sponza* and *sibenik* result in either brightening or darkening of corners. Darkening is subjectively acceptable, whereas brightening leads to objectionable artifacts.

metrics that performed the best for one data set (HDR-VDP and SSIM for NORM), ended up in the middle or the end of the ranking for the other data set (LOCCG). This is another example of the volatility of the comparison experiments, discussed in Section 4.2.4. Because of the large differences in metric performance between images, no metric could be said to be statistically significantly better (in terms of AUC) than any other metric in a general case. More helpful was the detailed analysis of the results for particular images, which revealed the issues that reduced the performance of the advanced metrics. One of those issues was excessive sensitivity to brightness and contrast changes, which are common in graphics due to the bias of rendering methods (refer to Figure 16). The simple metrics failed to distinguish between imperceptible and well visible noise levels in complex scenes (refer to Figure 17). The multi-scale metrics revealed problems in localizing small-area and high-contrast distortions (refer to Figure 18). But the most challenging are the distortions that appeared as a plausible part of the scene, such as darkening in corners, which appeared as soft shadows (refer to Figure 19).

Overall, the results revealed that the metrics are not as universal as they are believed to be. Complex metrics employing multi-scale decompositions can better predict visibility of low contrast distortions but they are less successful with suprathreshold distortions. Simple metrics, such as PSNR, can localize distortions well, but they fail to account for masking effects.

4.4.2 3D Model Quality Assessment

For model-based metrics (i.e. relying on the geometry), recent studies [44, 19] have provided extensive quantitative comparisons of existing metrics by computing correlations with MOS from several databases. Studies generally consider two correlation coefficients: the Spearman Rank Order Correlation Coefficient (SROC) which measures the monotonic association between the MOS and the metric values and the Pearson Linear Correlation Coefficient (LCC), which measures the prediction accuracy. The Pearson correlation is computed after performing a non-linear regression on the metric values as suggested by the video quality experts group (VQEG) [94], usually using a cumulative Gaussian function. Table 1 summarizes these correlation results; best metrics are highlighted for each database. Note that many metrics cannot be applied to evaluating simplification distortions because they need the compared objects to share the same connectivity – [32, 84, 45, 89, 43] – or the same level of details – [18].

Table 1 Correlation between Mean Opinion Scores and values from the metrics for four publicly-available subjective databases. The correlations are computed for whole databases, with the exception of the compression database, where per-model averages were used, since the data acquiring procedure does not capture inter-model relations.

	Masking [41]		Simplification [80]		General Purpose [45]		Compression [89]	
	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC
Hausdorff	0.27	0.20	0.49	0.51	0.14	0.11	0.25	0.14
RMS	0.49	0.41	0.64	0.59	0.27	0.28	0.52	0.49
GL1 [32]	0.42	0.40	NA	NA	0.33	0.35	0.67	0.71
GL2 [84]	0.40	0.38	NA	NA	0.39	0.42	0.74	0.76
3DWPM1 [18]	0.29	0.32	NA	NA	0.69	0.62	0.82	0.84
3DWPM2 [18]	0.37	0.43	NA	NA	0.49	0.50	0.81	0.82
MSDM [45]	0.65	0.69	NA	NA	0.74	0.75	**0.83**	**0.91**
MSDM2 [42]	**0.90**	**0.87**	**0.87**	**0.89**	**0.80**	**0.81**	**0.78**	**0.89**
FMPD [97]	0.80	0.81	**0.87**	**0.89**	**0.82**	**0.84**	**0.82**	**0.89**
DAME [89]	0.68	0.59	NA	NA	0.77	0.75	**0.86**	**0.94**
TPDM [86]	**0.87**	**0.87**	**0.87**	**0.86**	**0.85**	**0.85**	No data	No data
Learning [43]	**0.90**	**0.90**	NA	NA	**0.86**	**0.86**	No data	No data

We can observe that classical geometric distances, like Hausdorff and RMS, provide a very poor correlation with human judgement, while most recent ones [42, 97, 89, 86, 43] provide much better performance. Unfortunately, image-based metrics have not been quantitatively tested on these public databases, hence a legitimate question remains: which is the best to predict 3D mesh visual fidelity, image-based or model-based metrics? Rogowitz and Rushmeier [75] argue for model-based metrics since they show that 2D judgments do not provide a good predictor of 3D object quality, implying that the quality of 3D objects cannot be correctly predicted by the quality of static 2D projections. To demonstrate that, the authors have conducted two subjective rating experiments; in the first one, the observers

rated the quality of 2D static images of simplified 3D objects, while in the second one they rated an animated sequence of these images, showing a rotation of the 3D objects. Results show that (1) the lighting conditions strongly influence the perceived quality and (2) the observers perceive differently the quality of the 3D objects if they observe still images or animations. Watson et al. [101] also compared the performance of several image-based (Bolin-Meyer [12] and Mean Squared Error) and model-based (mean, max and RMS) metrics. They conducted several subjective experiments to study the visual fidelity of simplified 3D objects, including perceived quality rating. Their results showed a good performance of 2D metrics (Bolin-Meyer [12] and MSE) as well as the mean 3D geometric distance as predictor of the perceived quality. The main limitation of this study is that the authors only consider one single view of the 3D models. More recently, Cleju and Saupe [17] designed another subjective experiment for evaluating the perceived visual quality of simplified 3D models and found that generally image-based metrics perform better than model-based metrics. In particular, they found that 2D mean squared error and SSIM provide good results, whereas SSIM's performance being more sensitive to the 3D model type. For model-based metric, like Watson et al. [101], they showed that the mean geometric distance performs better than RMS which is better than Hausdorff (i.e. maximum distance). The main limitations of these studies (mostly from 10 years ago) is that they consider one single type of distortion (only simplification) and very simple image-based and model-based metrics. Lavoué et al. [46] have recently conducted a large experiment where they compare the most recent and efficient image metrics against the best model-based metrics. Their study considers a large set of parameters involved: several rendering algorithms, several lighting conditions and so on. Their conclusion is that image-based metrics are excellent for evaluating different versions of a same object under a same type of distortions, however they are less accurate at comparing different distortions or distortions applied on different 3D models. Hence, for simple use cases (e.g. determining the best parameters of a compression algorithm) image-based metrics will work very well; however, in scenarios involving the ranking of different distortions applied on different 3D objects (e.g. benchmarking different watermarking algorithms run on different sets of models) then model-based approaches should be preferred.

For dynamic meshes, a study presented by Váša and Skala [88] demonstrate an excellent prediction performance of the STED metric; while others (e.g. the KG error [33]) provide very poor results. Another open question concerns the quantitative evaluation of quality metrics for coloured or textured meshes; indeed per-vertex colours or texture play a very important role in the appearance of a 3D model however very few metrics still exist and no comparison study is still available.

5 Emerging Trends

5.1 Machine Learning

The objective of a quality assessment metric is to predict the visual quality of a signal, hence it basically needs to mimic the psychophysical process of the HSV, or at least relies on some features related to perceptual mechanisms. However modeling these complex principles and/or choosing appropriate characteristics may be hard. Hence it may appear convenient to treat the HVS as a black box which we wish to learn the input-output relationship. Such learning approaches were proposed recently [43, 90, 30]; they compute a large number of features and train classifiers on subjective ground truth data. Such kinds of metrics are usually very efficient however their ability to generalize depends on the richness of the ground-truth data. A very interesting point is that crowd-sourcing is developing as an excellent way to gather quickly a huge set of human opinions, that can then feed a classifier. As stated in the introduction, the future of quality metrics could lie in a combination of machine learning techniques with accurate psychophysical models.

5.2 3D animation

There still exist very few works about quality assessment for dynamic meshes (i.e. sequence of meshes) and articulated meshes (i.e. one single mesh + animated skeleton) while these types of data are present in a wide range of computer graphics applications. The perceived visual quality of such 3D animation depends not only on the geometry, texture and other visual attributes but also, to a large extent, on the nature of the movement and its velocity. This temporal dimension carries a whole range of additional cognitive phenomena. The contrast sensitivity function, for instance, is completely modified in a dynamic setting [34]. This is easily understandable since a rapid movement will be able to hide a geometrical artifact which would have been visible in a static case. In the case of human or animal animations, the *realism* of the animation is also a critical factor in the perception from the user. All these factors should be taken into account to devise efficient quality metrics, many progresses still remain to be achieved in this field.

5.3 Material and lighting

The need of photorealistic rendering of 3D content has led to embed complex material and lighting information together with the geometric information. For instance the Bi-directional Reflectance Distribution Function (BRDF) describes how much light is reflected when light makes contact with a certain material. More complex

non-uniform materials can be represented by more complex reflectance functions acquired through sophisticated photometric systems, including Surface Light Field (SLF) which represents the colour of a point depending on the viewing direction (hence assuming a fixed lighting direction), Bidirectional Texture Function (BTF) that extends the SLF for any incident lighting direction and finally Bidirectional Subsurface Scattering Reflectance Distribution Function (BSSRDF) which is basically a BTF plus a model of the surface scattering. There still exist no metric to assess the quality of these complex attributes (mapped or not onto the surface). In particular, it could be very useful to integrate them into existing model-based metrics (e.g. MSDM2) which are currently too much independent of the rendering conditions.

5.4 Toward merging image and model artefacts

We have seen all along this chapter that visual defects may appear at several stages of a computer graphics work-flow (as illustrated in figure 1) and may concern different types of data: either the 3D models, or the final rendered or tone mapped images. We have seen that there exist specific metrics dedicated to the detection of these model or image artifacts. Their use depends on the application, e.g. a 3D mesh compression approach has to be driven by a metric operating on the geometry, while a global illumination algorithm will be tuned using an image quality metric. What has been ignored until now is that these visual defects introduced either onto the geometry or onto the final images do have a visual interplay. For instance the nature of the rendering algorithm obviously influences the perceptibility of a geometric artefact; similarly, some types of rendering artefact could be avoided by a proper modelling or a specific geometry processing algorithm. Hence it appears obvious that these two types of quality assessment (i.e. respectively applied on models and images) should be connected. Integrating lighting and material information into model-based metrics (like mentioned in the above paragraph) could be a way to take into account these both processes (modeling and rendering). Considering the 3D scene for detecting image-based artefacts could be another way to model efficiently this interplay.

References

1. Akyüz, A.O., Fleming, R., Riecke, B.E., Reinhard, E., Bulthoff, H.H.: Do HDR displays support LDR content? A psychophysical evaluation. ACM Transactions on Graphics **26**(3), article no. 38 (2007)
2. Akyüz, A.O., Reinhard, E.: Perceptual evaluation of tone reproduction operators using the Cornsweet-Craik-O'Brien illusion. ACM Transactions on Applied Perception **4**(4), 1–29 (2008)

3. Allan, R., Terry, M.E.: Rank Analysis of Incomplete Block Designs : I . The Method of Paired Comparisons. Biometrica **39**(3/4), 324–345 (1952)
4. Ashikhmin, M., Goyal, J.: A reality check for tone mapping operators. ACM Transactions on Applied Perception **3**(4), 399–411 (2006)
5. Aydın, T.O., Mantiuk, R., Myszkowski, K., Seidel, H.P.: Dynamic range independent image quality assessment. ACM Transactions on Graphics (Proc. of SIGGRAPH) **27**(3), 69 (2008)
6. Aydın, T.O., Mantiuk, R., Seidel, H.P.: Extending quality metrics to full luminance range images. In: Proceedings of SPIE, pp. 68,060B–10. Spie (2008). DOI 10.1117/12.765095
7. Aydın, T.O., Čadík, M., Myszkowski, K., Seidel, H.P.: Video quality assessment for computer graphics applications. ACM Transactions on Graphics **29**(6), 1 (2010). DOI 10.1145/1882261.1866187
8. Banterle, F., Ledda, P., Debattista, K., Bloj, M., Artusi, A., Chalmers, A.: A Psychophysical Evaluation of Inverse Tone Mapping Techniques. Computer Graphics Forum **28**(1), 13–25 (2009). DOI 10.1111/j.1467-8659.2008.01176.x
9. Blackwell, H.: Contrast thresholds of the human eye. Journal of the Optical Society of America **36**(11), 624–632 (1946)
10. Boitard, R., Cozot, R., Thoreau, D., Bouatouch, K.: Temporal coherency in video tone mapping, a survey. In: HDRi2013 - First International Conference and SME Workshop on HDR imaging, Xx, p. no. 1 (2013)
11. Bolin, M.R., Meyer, G.W.: A frequency based ray tracer. In: Proc. of SIGGRAPH '95, pp. 409–418 (1995)
12. Bolin, M.R., Meyer, G.W.: A perceptually based adaptive sampling algorithm. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques - SIGGRAPH '98, pp. 299–309. ACM Press, New York, New York, USA (1998). DOI 10.1145/280814.280924
13. Bulbul, A., Capin, T., Lavoue, G., Preda, M.: Measuring Visual Quality of 3D Polygonal Models. IEEE Signal Processing Magazine **28**(6), 80–90 (2011)
14. Cater, K., Chalmers, A., Ward, G.: Detail to Attention: Exploiting Visual Tasks for Selective Rendering. Proc. of Eurographics workshop on Rendering pp. 270–280 (2003)
15. Cho, J., Prost, R., Jung, H.: An oblivious watermarking for 3-D polygonal meshes using distribution of vertex norms. IEEE Transactions on Signal Processing **55**(1), 142–155 (2007)
16. Cignoni, P., Rocchini, C., Scopigno, R.: Metro: Measuring Error on Simplified Surfaces. Computer Graphics Forum **17**(2), 167–174 (1998). DOI 10.1111/1467-8659.00236
17. Cleju, I., Saupe, D.: Evaluation of supra-threshold perceptual metrics for 3D models. In: Symposium on Applied Perception in Graphics and Visualization. ACM Press (2006). DOI 10.1145/1140491.1140499
18. Corsini, M., Gelasca, E.D., Ebrahimi, T., Barni, M.: Watermarked 3-D Mesh Quality Assessment. IEEE Transactions on Multimedia **9**(2), 247–256 (2007)
19. Corsini, M., Larabi, M.C., Lavoué, G., Petík, O., Váša, L., Wang, K.: Perceptual Metrics for Static and Dynamic Triangle Meshes. Computer Graphics Forum **32**(1), 101–125 (2013). DOI 10.1111/cgf.12001
20. Daly, S.: The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity. In: A.B. Watson (ed.) Digital Images and Human Vision, pp. 179–206. MIT Press (1993)
21. Delahunt, P.B., Zhang, X., Brainard, D.H.: Perceptual image quality: Effects of tone characteristics. Journal of Electronic Imaging **14**(2), 1–12 (2005). DOI 10.1117/1.1900134
22. Drago, F., L.Martens, W., Myszkowski, K., Sidel, H.P.: Perceptual Evaluation of Tone Mapping Operators with Regard to Similarity and Preference. Tech. rep., MPI Informatik (2002)
23. Dumont, R., Pellacini, F., Ferwerda, J.A.: Perceptually-driven decision theory for interactive realistic rendering. ACM Transactions on Graphics **22**(2), 152–181 (2003). DOI 10.1145/636886.636888
24. Eilertsen, G., Wanat, R., Mantiuk, R.K., Unger, J.: Evaluation of Tone Mapping Operators for HDR-Video. Computer Graphics Forum (Proc. of Pacific Graphics) **32**(7), 275–284 (2013)
25. Engeldrum, P.: Psychometric scaling: a toolkit for imaging systems development. Imcotek Press (2000)

26. Ferwerda, J.A., Shirley, P., Pattanaik, S.N., Greenberg, D.P.: A model of visual masking for computer graphics. In: Proc. of SIGGRAPH '97, pp. 143–152. ACM Press, New York, New York, USA (1997). DOI 10.1145/258734.258818
27. Georgeson, M.A., Sullivan, G.D.: Contrast constancy: deblurring in human vision by spatial frequency channels. J. Physiol. **252**(3), 627–656 (1975)
28. Guthe, M., Müller, G., Schneider, M., Klein, R.: BTF-CIELab: A Perceptual Difference Measure for Quality Assessment and Compression of BTFs. Computer Graphics Forum **28**(1), 101–113 (2009). DOI 10.1111/j.1467-8659.2008.01299.x
29. Hadziabdic, K.K., Telalovic, J.H., Mantiuk, R.: Comparison of deghosting algorithms for multi-exposure high dynamic range imaging. In: Proc. of Spring Conference on Computer Graphics, pp. 1–8 (2013)
30. Herzog, R., Čadík, M., Aydın, T.O., Kim, K.I., Myszkowski, K., Seidel, H.P.: NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis. Computer Graphics Forum **31**(2pt3), 545–554 (2012). DOI 10.1111/j.1467-8659.2012.03055.x
31. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998). DOI 10.1109/34.730558
32. Karni, Z., Gotsman, C.: Spectral compression of mesh geometry. In: ACM Siggraph, pp. 279–286 (2000)
33. Karni, Z., Gotsman, C.: Compression of Soft-Body animation sequences. Computers & Graphics **28**(1), 25–34 (2004)
34. Kelly, D.H.: Motion and vision. {II}. Stabilized spatiotemporal threshold surface. Journal of Optical Society of America **69**(10), 1340–1349 (1979)
35. Kim, K.J., Mantiuk, R., Lee, K.H.: Measurements of achromatic and chromatic contrast sensitivity functions for an extended range of adaptation luminance. In: B.E. Rogowitz, T.N. Pappas, H. de Ridder (eds.) Human Vision and Electronic Imaging, p. 86511A (2013). DOI 10.1117/12.2002178
36. Korshunov, P., Ebrahimi, T.: Influence of Context and Content on Tone-mapping Operators. In: HDRi2013 - First International Conference and SME Workshop on HDR imaging, p. no. 2 (2013)
37. Krawczyk, G., Myszkowski, K., Seidel, H.P.: Contrast Restoration by Adaptive Countershading. Computer Graphics Forum **26**(3), 581–590 (2007). DOI 10.1111/j.1467-8659.2007.01081.x
38. Kuang, J., Heckaman, R., Fairchild, M.D.: Evaluation of HDR tone-mapping algorithms using a high-dynamic-range display to emulate real scenes. Journal of the Society for Information Display **18**(7), 461–468 (2010). DOI 10.1889/JSID18.7.461
39. Kuang, J., Yamaguchi, H., Johnson, G.M., Fairchild, M.D.: Testing HDR image rendering algorithms. In: Proc. IS&T/SID 12th Color Imaging Conference, pp. 315–320. Scotsdale, Arizona (2004)
40. Larkin, M., O'Sullivan, C.: Perception of Simplification Artifacts for Animated Characters. In: symposium on Applied perception in graphics and visualization, pp. 93–100 (2011)
41. Lavoué, G.: A local roughness measure for 3D meshes and its application to visual masking. ACM Transactions on Applied Perception (TAP) **5**(4) (2009)
42. Lavoué, G.: A Multiscale Metric for 3D Mesh Visual Quality Assessment. Computer Graphics Forum **30**(5), 1427–1437 (2011)
43. Lavoué, G., Cheng, I., Basu, A.: Perceptual Quality Metrics for 3D Meshes: Towards an Optimal Multi-Attribute Computational Model. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2013)
44. Lavoué, G., Corsini, M.: A comparison of perceptually-based metrics for objective evaluation of geometry processing. IEEE Transactions on Multimedia **12**(7), 636–649 (2010)
45. Lavoue, G., Drelie Gelasca, E., Dupont, F., Baskurt, A., Ebrahimi, T.: Perceptually driven 3D distance metrics with application to watermarking. In: SPIE, vol. 6312, pp. 63,120L–63,120L–12. SPIE (2006)

46. Lavoué, G., Larabi, M.C., Váša, L.: On the Efficiency of Image Metrics for Evaluating the Visual Quality of 3D Models. Submitted to IEEE Transactions on Visualization and Computer Graphics (2013)
47. Ledda, P., Chalmers, A., Troscianko, T., Seetzen, H.: Evaluation of tone mapping operators using a high dynamic range display. ACM Transactions on Graphics **24**(3), 640–648 (2005)
48. Lindstrom, P.: Model Simplification using Image and Geometry-Based Metrics. Ph.D. thesis, Georgia Institute of Technology (2000)
49. Lindstrom, P., Turk, G.: Evaluation of memoryless simplification. IEEE Transactions on Visualization and Computer Graphics **5**(2), 98–115 (1999). DOI 10.1109/2945.773803
50. Lindstrom, P., Turk, G.: Image Driven Simplification. ACM Transactions on Graphics **19**(3), 204–241 (2000)
51. Liu, Y., Wang, J., Cho, S., Finkelstein, A., Rusinkiewicz, S.: A no-reference metric for evaluating the quality of motion deblurring. ACM Transactions on Graphics **32**(6), 1–12 (2013). DOI 10.1145/2508363.2508391
52. Lubin, J.: A visual discrimination model for imaging system design and evaluation. In: E. Peli (ed.) Vision Models for Target Detection and Recognition, pp. 245–283. World Scientific Publishing Company (1995)
53. Luebke, D., Hallen, B.: Perceptually driven simplification for interactive rendering. In: Rendering Techniques 2001: Proceedings of the Eurographics Workshop, p. 223 (2001)
54. Luebke, D., Hallen, B., Newfield, D., Watson, B.: Perceptually Driven Simplification Using Gaze-Directed Rendering. In: EGSR, pp. 223–234 (2001)
55. Mantiuk, R., Daly, S., Myszkowski, K., Seidel, H.: Predicting visible differences in high dynamic range images: model and its calibration. In: Human Vision and Electronic Imaging, pp. 204–214 (2005)
56. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graph (Proc. SIGGRAPH) **30**(4), 1 (2011). DOI 10.1145/2010324.1964935
57. Mantiuk, R.K., Tomaszewska, A., Mantiuk, R.: Comparison of four subjective methods for image quality assessment. Computer Graphics Forum **31**(8), 2478–2491 (2012)
58. Masia, B., Agustin, S., Fleming, R.W., Sorkine, O., Gutierrez, D.: Evaluation of reverse tone mapping through varying exposure conditions. ACM Transactions on Graphics **28**(5), 1 (2009). DOI 10.1145/1618452.1618506
59. McCann, J., Rizzi, A.: Veiling glare: the dynamic range limit of hdr images. In: Proc. of HVEI XII, vol. 6492, pp. 649,213–649,213. International Society for Optics and Photonics (2007)
60. McCann, J.J., Rizzi, A.: The Art and Science of HDR Imaging(Google eBook). John Wiley & Sons (2011)
61. Menzel, N., Guthe, M.: Towards Perceptual Simplification of Models with Arbitrary Materials. Computer Graphics Forum **29**(7), 2261–2270 (2010). DOI 10.1111/j.1467-8659.2010.01815.x
62. Mullen, K.T.: The contrast sensitivity of human colour vision to red-green and blue-yellow chromaic gratings. Journal of Physiolohy **359**, 381–400 (1985)
63. Myszkowski, K.: The visible differences predictor: Applications to global illumination problems. In: Rendering techniques' 98: proceedings of the Eurographics Workshop in Vienna, Austria, June 29-July 1, 1998, p. 223. Springer Verlag Wien (1998)
64. Myszkowski, K., Rokita, P., Tawara, T.: Perceptually-informed accelerated rendering of high quality walkthrough sequences. In: Eurographics Workshop on Rendering, vol. 99, pp. 5–18 (1999)
65. Myszkowski, K., Tawara, T., Akamine, H., Seidel, H.P.: Perception-guided global illumination solution for animation rendering. In: Proc. of SIGGRAPH'01, pp. 221–230. ACM, New York, NY, USA (2001). DOI 10.1145/383259.383284
66. O'Donovan, P., Agarwala, A., Hertzmann, A.: Color compatibility from large datasets. ACM Transactions on Graphics **30**(4), 1 (2011). DOI 10.1145/2010324.1964958
67. O'Sullivan, C., Howlett, S., Morvan, Y.: Perceptually adaptive graphics. Eurographics State of the Art Reports pp. 141 – 164 (2004)

68. Pan, Y., Cheng, I., Basu, A.: Quality metric for approximating subjective evaluation of 3-D objects. IEEE Transactions on Multimedia **7**(2), 269–279 (2005)
69. Petit, J., Mantiuk, R.K.: Assessment of video tone-mapping : Are cameras S-shaped tone-curves good enough? Journal of Visual Communication and Image Representation **24**, 1020–1030 (2013). DOI 10.1016/j.jvcir.2013.06.014
70. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: TID2008 - A database for evaluation of full-reference visual quality assessment metrics. Advances of Modern Radioelectronics **10**, 30–45 (2009)
71. Qu, L., Meyer, G.: Perceptually guided polygon reduction. IEEE Transactions on Visualization and Computer Graphics **14**(5), 1015–1029 (2008). DOI 10.1109/TVCG.2008.51
72. Ramanarayanan, G., Ferwerda, J., Walter, B.: Visual equivalence: towards a new standard for image fidelity. ACM Transactions on Graphics (TOG) **26**(3), 76 (2007). DOI 10.1145/1276377.1276472
73. Ramasubramanian, M., Pattanaik, S.N., Greenberg, D.P.: A perceptually based physical error metric for realistic image synthesis. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99, pp. 73–82. ACM Press, New York, New York, USA (1999). DOI 10.1145/311535.311543
74. Reddy, M.: SCROOGE: Perceptually-Driven Polygon Reduction. Computer Graphics Forum **15**(4), 191–203 (1996)
75. Rogowitz, B.E., Holly E. Rushmeier: Are image quality metrics adequate to evaluate the quality of geometric objects? Proceedings of SPIE pp. 340–348 (2001)
76. Rubinstein, M., Gutierrez, D., Sorkine, O., Shamir, A.: A comparative study of image retargeting. ACM Transactions on Graphics **29**(6), 1 (2010). DOI 10.1145/1882261.1866186
77. Rushmeier, H., Rogowitz, B., Piatko, C.: Perceptual issues in substituting texture for geometry. In: SPIE, pp. 372–383. International Society for Optical Engineering; 1999 (2000)
78. Schroeder, W., Zarge, J., Lorensen, W.: Decimation of triangle meshes. In: ACM Siggraph, pp. 65–70 (1992)
79. Secord, A., Lu, J., Finkelstein, A., Singh, M., Nealen, A.: Perceptual models of viewpoint preference. ACM Transactions on Graphics **30**(5), 1–12 (2011). DOI 10.1145/2019627.2019628
80. Silva, S., Santos, B., Ferreira, C.: Comparison of methods for the simplification of mesh models using quality indices and an observer study. SPIE pp. 64,921L–64,921L–12 (2007)
81. Silva, S., Santos, B.S., Ferreira, C., Madeira, J.: A Perceptual Data Repository for Polygonal Meshes. 2009 Second International Conference in Visualisation pp. 207–212 (2009)
82. Silverstein, D., Farrell, J.: Efficient method for paired comparison. Journal of Electronic Imaging **10**, 394 (2001). DOI 10.1117/1.1344187
83. Smith, K., Krawczyk, G., Myszkowski, K.: Beyond tone mapping: Enhanced depiction of tone mapped HDR images. Computer Graphics Forum **25**(3), 427–438 (2006)
84. Sorkine, O., Cohen-Or, D., Toldeo, S.: High-pass quantization for mesh encoding. In: Eurographics Symposium on Geometry Processing, pp. 42–51 (2003)
85. Tian, D., AlRegib, G.: FQM: A Fast Quality Measure for Efficient Transmission of Textured 3D Models. In: ACM Multimedia, pp. 684–691 (2004)
86. Torkhani, F., Wang, K., Chassery, J.m.: A Curvature Tensor Distance for Mesh Visual Quality Assessment. In: International Conference on Computer Vision and Graphics (2012)
87. Trentacoste, M., Mantiuk, R., Heidrich, W., Dufrot, F.: Unsharp Masking, Countershading and Halos: Enhancements or Artifacts? Computer Graphics Forum **31**(2pt3), 555–564 (2012). DOI 10.1111/j.1467-8659.2012.03056.x
88. Vasa, L., Skala, V.: A Perception Correlated Comparison Method for Dynamic Meshes. IEEE Trans. on Visualization and Computer Graphics **17**(2), 220–230 (2011)
89. Váša, L., Rus, J.: Dihedral Angle Mesh Error: a fast perception correlated distortion measure for fixed connectivity triangle meshes. Computer Graphics Forum **31**(5) (2012)
90. Čadík, M., Herzog, R., Mantiuk, R., Mantiuk, R., Myszkowski, K., Seidel, H.P.: Learning to Predict Localized Distortions in Rendered Images. Computer Graphics Forum (Proc. of Pacific Graphics) **32**(7), 401–410 (2013)

91. Čadík, M., Herzog, R., Mantiuk, R.K., Myszkowski, K., Seidel, H.P., Čadík, M.: New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts. ACM Trans. Graph (Proc. SIGGRAPH Asia) **31**(6), 147 (2012). DOI 10.1145/2366145.2366166
92. Čadík, M., Wimmer, M., Neumann, L., Artusi, A.: Evaluation of HDR tone mapping methods using essential perceptual attributes. Computers & Graphics **32**(3), 330–349 (2008). DOI 10.1016/j.cag.2008.04.003
93. Villa, C., Labayrade, R.: Psychovisual assessment of tone-mapping operators for global appearance and colour reproduction. In: Proc. of Colour in Graphics Imaging and Vision 2010, pp. 189–196. Joensuu, Finland (2010)
94. VQEG: Final report from the video quality experts group on the validation of objective models of video quality assessment. Tech. rep., Video Quality Experts Group (2000)
95. Walter, B., Pattanaik, S.N., Greenberg, D.P.: Using Perceptual Texture Masking for Efficient Image Synthesis. Computer Graphics Forum **21**(3), 393–399 (2002). DOI 10.1111/1467-8659.t01-1-00599
96. Wang, K., Lavoué, G., Denis, F., Baskurt, A.: Robust and blind mesh watermarking based on volume moments. Computers & Graphics **35**(1), 1–19 (2011)
97. Wang, K., Torkhani, F., Montanvert, A.: A Fast Roughness-Based Approach to the Assessment of 3D Mesh Visual Quality. Computers & Graphics (2012)
98. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
99. Watson, A., Ahumada Jr, A.: A standard model for foveal detection of spatial contrast. Journal of Vision **5**(9), 717–740 (2005)
100. Watson, A.B.: The cortex transform: Rapid computation of simulated neural images. Computer Vision, Graphics, and Image Processing **39**(3), 311–327 (1987). DOI 10.1016/S0734-189X(87)80184-6
101. Watson, B., Friedman, A., McGaffey, A.: Measuring and predicting visual fidelity. ACM Siggraph pp. 213–220 (2001)
102. Williams, N., Luebke, D., Cohen, J., Kelley, M., Schubert, B.: Perceptually Guided Simplification of Lit, Textured Meshes. In: ACM Symposium on Interactive 3D Graphics, pp. 113–121 (2003)
103. Wilson, H.R.: A transducer function for threshold and suprathreshold human vision. Biological Cybernetics **38**(3), 171–178 (1980). DOI 10.1007/BF00337406
104. Yee, H.: Perceptual Metric for Production Testing. Journal of Graphics Tools **9**(4), pages 33–40 (2004)
105. Yee, H., Pattanaik, S., Greenberg, D.P.: Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. ACM Transactions on Graphics **20**(1), 39–65 (2001). DOI 10.1145/383745.383748
106. Yeganeh, H., Wang, Z.: Objective quality assessment of tone-mapped images. IEEE Transactions on Image Processing **22**(2), 657–67 (2013). DOI 10.1109/TIP.2012.2221725
107. Yoshida, A., Blanz, V., Myszkowski, K., Seidel, H.P.: Perceptual evaluation of tone mapping operators with real world scenes. In: Proc. of SPIE Human Vision and Electronic Imaging X, vol. 5666, pp. 192–203. San Jose, CA (2005)
108. Yoshida, A., Mantiuk, R., Myszkowski, K., Seidel, H.P.: Analysis of reproducing real-world appearance on displays of varying dynamic range. Computer Graphics Forum **25**(3), 415–426 (2006)
109. Zhang, X., Wandell, B.A.: A spatial extension of CIELAB for digital color-image reproduction. Journal of the Society for Information Display **5**(1), 61 (1997). DOI 10.1889/1.1985127
110. Zhu, Q., Zhao, J., Du, Z., Zhang, Y.: Quantitative analysis of discrete 3D geometrical detail levels based on perceptual metric. Computers & Graphics **34**(1), 55–65 (2010). DOI 10.1016/j.cag.2009.10.004