Quantifying image quality in graphics: Perspective on subjective and objective metrics and their performance

Rafał K. Mantiuk

Bangor University, School of Computer Science, Bangor, UK

ABSTRACT

We explore three problems related to quality assessment in computer graphics: the design of efficient user studies; the scene-referred metrics for comparing high-dynamic-range images; and the comparison of metric performance for the database of computer graphics distortions. This paper summarizes the most important observations from investigating these problems and gives a high level perspective on the problem of quality assessment in graphics.

1. SUMMARY

As the results of computer graphics methods are typically visual, image quality assessment is one of its central problems. The problem is similar to quality assessment in video compression and transmission, but there are also several important differences. These differences motivate the need for more sensitive and time-efficient subjective metrics (quality assessment experiments), objective metrics that operate on scene-referred rather than display-referred images, and image quality assessment methods that give accurate predictions on the per-pixel rather than per-image basis.

In this paper we first identify the problems that are unique to graphics and review available solutions (Section 1.1). Then we discuss in more detail the three problems that we have investigated: time-efficiency of subjective quality assessment methods (Section 1.2); objective quality metrics for high dynamic range (HDR) images (Section 1.3); and the performance of popular quality metrics for the data-base of computer graphics distortions (Section 2). This paper summarizes the work that has been explored in more detail elsewhere,[1–3] and presents a high level perspective on the problem of quality assessment in graphics (Section 3).

1.1 Image quality assessment problems in graphics

Quality assessment in graphics poses several challenges that make it quite distinct from other image quality applications. We review them in the following paragraphs.

High dimensionality of possible result space. Many computer graphics methods involve a large number of adjustable parameters, each producing a slightly different resulting image. One example of that is tone-mapping, where the results can be significantly improved if the parameters are adjusted individually per-image. The question then arises of how to explore the space of possible results and to find the best image to compare against competing methods. Even if the authors can find the best set of parameters for their method, they will be less inclined to spend an equal amount of time fine-tuning competitive methods for a fair comparison. Therefore, a number of subjective tone-mapping evaluation studies could be biased because the results produced by each method may not have been selected in a fair manner.

The obvious solution is to explore the entire space of possible parameter values and solutions. In the case of video compression, this is a straightforward procedure which requires assessing quality at several bit-rates and plotting quality vs. bit-rate curves for each method. However, if a method involves two or more parameters, brute-force quality assessment is impractical, especially if subjective methods need to be used. Objective methods could be used instead, but then an appropriate metric must be available (and there is little choice for automatic tone-mapping evaluation) and must prove its reliability.

No reference. The goal of computer rendering is to generate an image from a geometric definition, without knowing what the final image should be. Therefore, in rendering applications a reference image is usually not

Further author information: (Send correspondence to R.K.M)
R.K.M.: E-mail: mantiuk@bangor.ac.uk, Telephone: 0044 12483 82689

available and only the geometric definition of a scene is given. Herzog et al.[4] proposed a non-reference metric for graphics, which used the information about the underlying synthetic scene (e.g., 3D surfaces, textures) instead of considering color alone in order to train a classifier. The metric demonstrated very promising results for the training test set consisting of 10 manually marked images. The robustness of such a non-parametric classifiers still needs to be tested on a larger data set.

On-line application of the metrics. Many graphics or image processing algorithms are formulated as an optimization problem in which the difference between an ideal and a feasible solution is minimized. Such feasible solution may impose constraints in terms of computation time, available color gamut, or similar. Although the least-squares are the most common criteria used for optimization, it can be argued that better results can be achieved if a visual metric is used instead. Such an approach was proposed for gamut-mapping[5] and tone-mapping.[6] Visual metrics were also used to drive the solution of computer graphics rendering methods.[7–9] These methods do not have a reference "ideal" image available, as discussed above, but this problem is remedied by using intermediate rendering results and their error estimate,[7] consecutive animation frames[8] or by comparing the results of two intermediate iterations of a rendering method.[9]

Invariance to non-relevant differences. Many graphics methods often attempt to produce plausible, rather than accurate images. They are designed to take advantage of the fact that human observers are not very sensitive to overall contrast and brightness changes, smoothing which does not blur edges, small geometric distortions, changes in shading or illumination, etc. For example, it is unlikely that an observer will notice that an object is shifted by a few pixels relative to a reference. However, the majority of pixel-based quality metrics will report very large errors for all misaligned pixels. Another example are brightness and contrast changes, which are very difficult to notice for most observers. They may be the result of a bias of graphics rendering methods. Biased methods often produce plausible and good looking images in shorter time, but they do not need converge to the physically accurate solution. Since such a bias is inconsistent through an image, most objective metrics will mark these inaccuracies as artifacts when compared to the physically-accurate reference. All these cases demonstrate that a visible distortion does not need to be seen as objectionable and a less conservative error estimate is likely to better correlate with subjective quality assessment.

The concept of *visual equivalence* was proposed[10] to distinguish between visible and objectionable image distortions. 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. Such definition leads to an experimental method, however, there is no objective metric that could predict equivalence in a general case.

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). In both cases, smooth gradients are introduced at both sides of an edge in order to enhance the contrast of that edge. It is possible to find a gradient that is invisible in an image but still introduces contrast enhancement (Cornsweet illusion). However, such enhancement is very small, impractical in most cases. A useful contrast enhancement requires introducing a visible gradient. But too strong gradient results in visible contrast reversal, also known as "halo" artifact. Figure 1 illustrates that work[11] that has been done on the problem of finding the threshold of just-objectionable contrast enhancement in order to better control it in tone-mapping and other applications.

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.[12] 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.[13] proposed a metric for comparing HDR and tone-mapped images that is robust to contrast changes. The metric was later extended to video.[14] 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

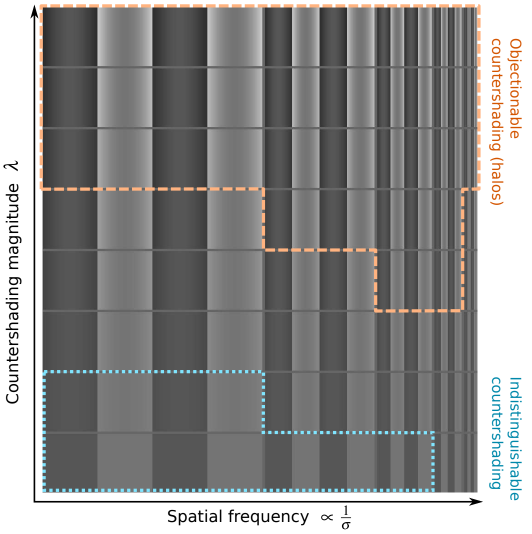


Figure 1: The square-wave pattern with a reduced amplitude of the fundamental frequency, resulting in countershading profiles. The regions of indistinguishable (from a step edge) and objectionable countershading are marked with dotted and dashed lines of different color. The higher magnitude of countershading 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. Note that the spatial frequency shown on this plot is inversely proportional to the profile width σ, which is used in the rest of the paper.

and contrast reversal. These three cases are believed to affect the quality of 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.

Quality evaluation for stereo images and disparity. Color is probably the most important, but not the only visual stimulus that lets us perceive 3D scenes. Binocular disparity is an important cue that has a strong influence on our depth perception. The introduction of stereo 3D technologies brought a broad range of applications which need to manipulate depth and disparity. Once disparity is manipulated, the question then arises of how visible are those changes to the human observer. Didyk et al.[15] proposed a perceptual model for disparity. Their method decomposes input disparity maps into three frequency bands and employs transducers to transform band-limited disparities into JND units. The transducers are derived from the disparity discrimination thresholds measured in a dedicated experiment. Once the disparities are represented in the JND units, they can be compared by taking a difference and employing the Minkowski summation to find a single-valued perceived difference estimate. In the following paper[16] the metric was further extended to account for the effect of luminance on the perception of disparity. Another aspect of stereo vision is binocular-rivalry, which is the conflict evoked when two different images are show to both eyes. Yang et al.[17] proposed a model predicting acceptable level of rivalry and applied it to tone-mapping, where different images are presented to both eyes in order to enhance contrast perception.

Quality evaluation for high dynamic range images. 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.

Aydin et al.[18] proposed a simple luminance encoding that makes it possible to use PSNR and SSIM[19] metrics with HDR images. The encoding transform physical luminance values (represented in cd/m^2) into an approximately perceptually uniform representation. The transformation is derived from luminance detection data using Fechnerian integration. 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 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 the most important luminance-dependent effect, which is the reduced sensitivity at lower luminance levels, but it does not account for other luminance 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).[20] The HDR-VDP extends Daly's visual difference predictor (VDP)[21] to predict differences in high dynamic range images. The second revision of the metric, HDR-VDP-2,[2] replaced most visual model components with better alternatives, introduced cross-channel contrast masking, flattening of the CSF at super-threshold contrast, and much improved the accuracy, especially at low luminance levels. We discuss HDR-VDP-2 in more detail in Section 1.3.

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 dataset 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,[22] or selecting color palettes for graphic designs.[23] 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 the topic of the next section.

1.2 Pairwise comparisons excel in sensitivity and time-efficiency

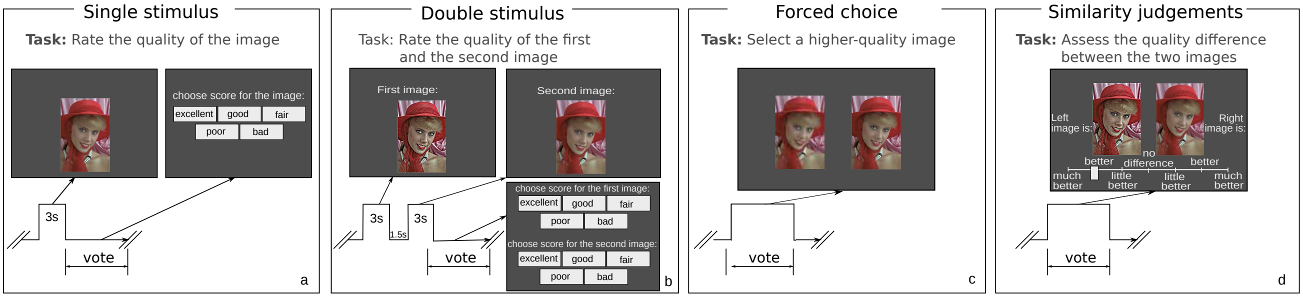


Figure 2: Overview of the four subjective quality assessment methods, which we compared in terms of sensitivity and observer's time-effort. The diagram shows the time-line of each method and the corresponding screens.

It is a common expectation that the results of a computer graphics or imaging algorithm are rigorously validated, preferably in a user study. The method of pairwise comparisons tends to be the most often used in graphics. Intuitively, it could be expected that pairwise comparison method, where the observer has to choose a better of two images, is easier for observers, easier to reproduce, and thus more accurate than direct rating, where the observer needs to assign a numerical value to each image individually. This intuition was confirmed in a formal study,[1] where four subjective methods of quality assessment were compared: single and double stimulus methods, forced choice pairwise comparison and similarity judgements. All these methods are visually illustrated in Figure 2. In case of both the force-choice and similarity judgements methods, the number of comparisons was reduced using an efficient sorting algorithm.[24] Refer to the paper[1] for more details on the experimental setup.

The sensitivity of each method was measured in terms of the effect size d, which is defined as the difference between a pair of quality scores normalized by a common standard deviation. The values of effect size for each method are shown in Figure 3. The forced choice pair-wise comparison method results in statistically significantly higher effect size as compared to both single and double stimulus methods. But the difference is not very large

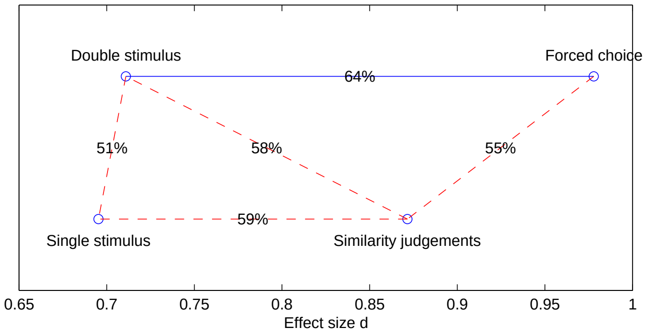


Figure 3: The comparison of effect size for each experimental method. The larger the effect size, the more accurate the method is. The y-axis is used only to better layout the methods and show their relations. The percentages indicate the probability that for a random pair of scenes and distortion types, the method on the right will result in higher sensitivity than the method on the left. If the line connecting two conditions is red and dashed, it indicates that there is no statistical difference between this pair of conditions (H_o could not be rejected for $\alpha = 0.05$ and adjusted for multiple comparisons).

in practical terms; only in about 64% of cases the forced choice method will result higher sensitivity than the double stimulus method, assuming the same sample size. Therefore, our data shows that the difference between direct rating and pairwise comparison methods exists, but it does not seem to be as dramatic as the four-fold reduction of standard deviation reported in.[25]

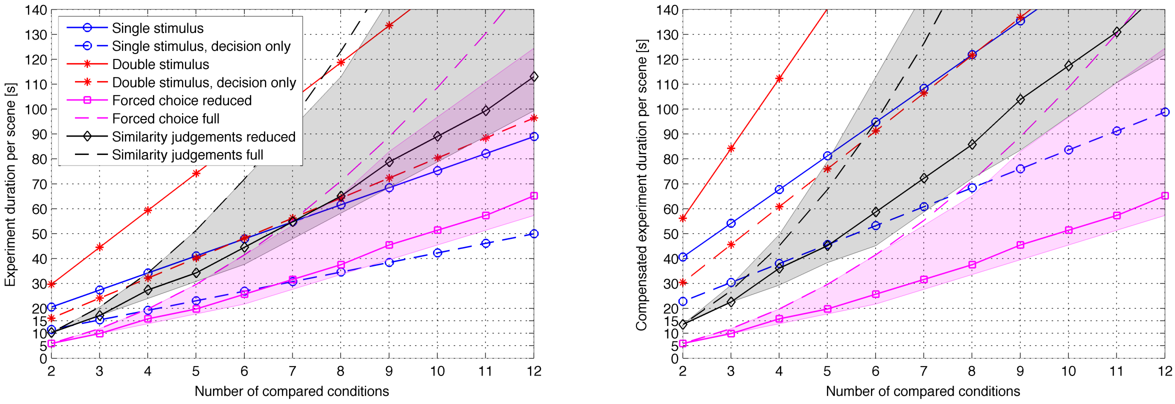
Even though the pair-wise comparison methods are marginally more sensitive, they have also the reputation of being tedious and requiring a very large number of trials. However, we found that when a reduced pairwise comparison design is used,[24] pairwise comparison method can be significantly faster than direct rating methods, even for a large number of compared conditions.[1] The results of that finding are shown in Figure 4a, where we compared the time required to compare a given set of conditions (algorithms and parameter variations) for a single image and a single trial. Since the experimental methods differ in their sensitivity, some methods may require more measurements to result in the same confidence intervals as the other methods. To account for this difference, Figure 4b shows the times compensated for the difference in the effect size. The time for worse performing methods was increased relative to the most accurate method — the forced choice pairwise comparison. After compensating the times, it is clear that both rating methods are significantly less effective than the pairwise comparison methods, especially the forced-choice method. If the reduced design is used, the single stimulus method does not seem to be more effective even if a large number of conditions is considered.

1.3 Visibility metric for high dynamic range images

Even though subjective methods, discussed in the previous section, will remain the most robust and reliable practice for assessing image quality, there is a number of applications where a suitable computational (objective) metric is essential. This section summarizes such an objective metric, which builds upon a vast body of research on detection and discrimination performance of the visual system. The metric is also an effort to design a comprehensive model of the contrast visibility for a very wide range of illumination conditions. It focuses on proper modelling of luminance-dependent effects and visual masking, both relevant for graphics.

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.[2] The metric originates from the classical Visual Difference Predictor,[21] and its extension — HDR-VDP.[20] 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.

The multiple stages of visual processing are described in detail elsewhere.[2] Here, we summarize the most important parts. One of the major factors limiting our contrast perception in high contrast (HDR) scenes is the scattering of the light in the optics of the eye and on the retina.[27] 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). Our contrast



(a) Non-compensated time

(b) Compensated time

Figure 4: (a) — time required to compare a given number of conditions (x-axis) using each experimental method. (b) — the same time that is compensated to result in the same relative width of the confidence intervals. The plots are based on the average time recorded in our experiments. Because the number of trials in reduced pairwise methods depends on the complexity of a sorting algorithm, the shaded regions represent the bounds between the best- and worst-case scenario. The continuous lines indicate the times based on the average complexity. The times include the assessment of a reference image for all methods, i.e. 2 conditions point corresponds to the assessment of two test images and one reference image. Non-smooth shape is due to rounding to an integer number of comparisons.

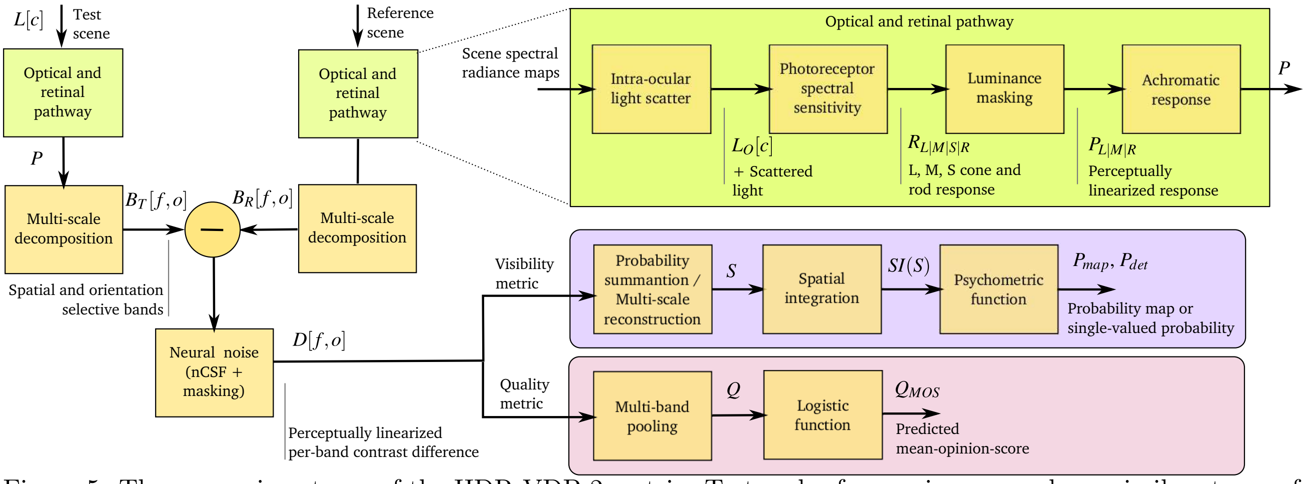
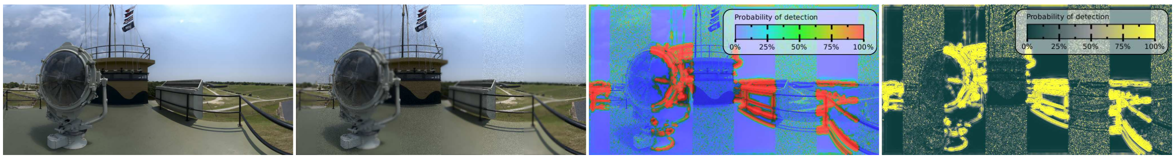

$L[c]$
Test scene
Reference scene
Optical and retinal pathway
Optical and retinal pathway
P
Multi-scale decomposition
Multi-scale decomposition
$B_T[f,o]$
$B_R[f,o]$
Spatial and orientation selective bands
Neural noise (nCSF + masking)
$D[f,o]$
Perceptually linearized per-band contrast difference
Scene spectral radiance maps
Optical and retinal pathway
Intra-ocular light scatter
Photoreceptor spectral sensitivity
Luminance masking
Achromatic response
P
$L_O[c]$
+ Scattered light
$R_{L|M|S|R}$
L, M, S cone and rod response
$P_{L|M|R}$
Perceptually linearized response
Visibility metric
Probability summantion / Multi-scale reconstruction
S
Spatial integration
$SI(S)$
Psychometric function
P_{map}, P_{det}
Probability map or single-valued probability
Quality metric
Multi-band pooling
Q
Logistic function
Q_{MOS}
Predicted mean-opinion-score

Figure 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).

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 comprehensive masking models (*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



Reference Image
Test image
Probability of detection (screen, color)
Probability of detection (print, dichromatic)

Figure 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.

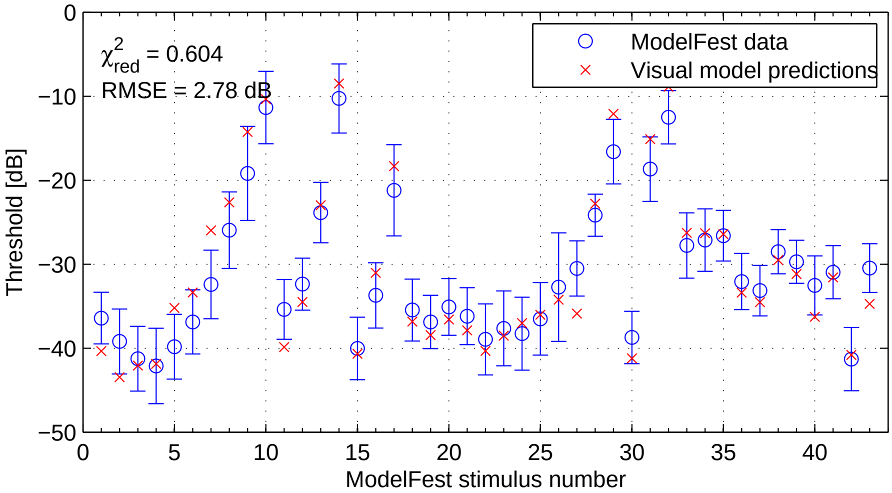


Figure 7: Visual model predictions for the **ModelFest** data set. Error bars denote standard deviation of the measurements. The R value is the prediction mean square root error and χ^2_{red} is the reduced chi-square statistic.

effect of neural CSF, which is the contrast sensitivity function without the sensitivity reduction due to interocular 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.[28]

Figure 6 demonstrates the metric prediction for blur and noise. The model has been thoroughly calibrated and shown to predict numerous discrimination data sets, such as ModelFest[29] (Figure 7), historical Blackwell's t.v.i. measurements[26] (Figure 8), and newly measured CSF (Figure 9). The source code of the metric is freely available for download from `http://hdrvdp.sourceforge.net` and the calibration data sets are available on request. It is also possible to run the metric using an on-line web service at `http://driiqm.mpi-inf.mpg.de/`.

2. PERFORMANCE OF QUALITY METRICS FOR COMPUTER GRAPHICS DISTORTIONS

It could be expected that graphics community could take advantage of the large number of objective quality metrics to assess quality degradation in computer generated images. We found, however, that the performance of the popular quality metrics is very inconsistent when tested with the distortions generated by computer graphics methods. In general case, no metric was found statistically significantly better than any other tested metric, even PSNR.

To compare the performance of popular quality metrics in graphics applications, we measured the visibility of graphics distortions by asking the observers to directly mark them in images using a brush-painting interface.[3] This is in contrast to typical subjective quality assessment methods, which assign a single quality value per

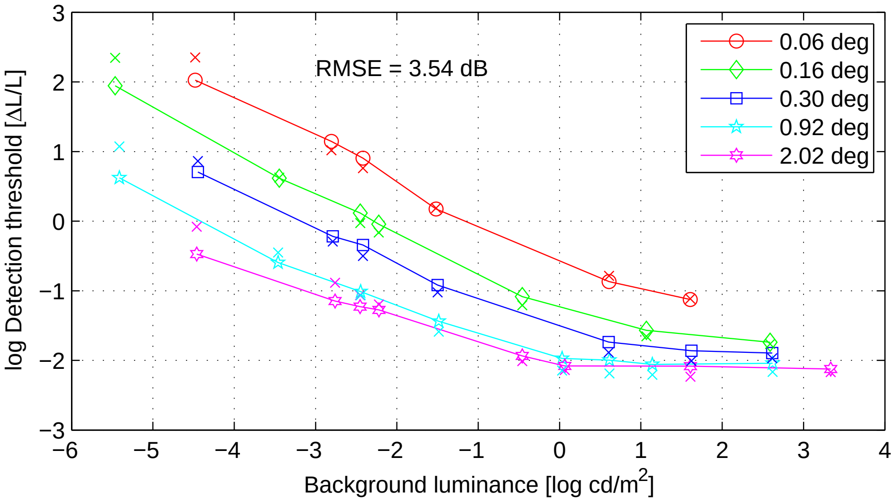


Figure 8: Visual model predictions for the **threshold versus intensity curve** data set. These are the detection thresholds for circular patterns of varying size on a uniform background field.[26]

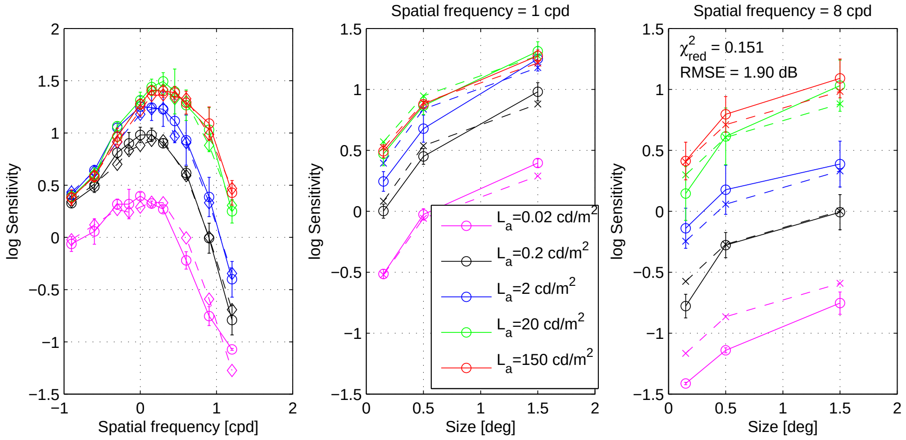


Figure 9: Visual model predictions for the **CSF for wide luminance range** data set. The two plots on the right show sensitivity variation due to stimuli size.

image.[25,30] The information about localization of distortions is very important in many graphics applications, in which a higher computational effort is directed towards more distorted regions. Such per-pixel image-quality data set is definitely more challenging for the quality metrics, as it requires producing consistent distortion maps, rather than fitting one quality value per image.

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 quite consistent across observers resulting in similar distortion maps for each individual. Despite the lack of reference, the observers could spot most of the artifacts in the first task. There was only marginal difference in terms of which image regions were marked in both tasks. The main difference was that fewer people could spot the artifacts when no-reference was provided. These results lead to two observations. Firstly, human observers are able to spot objectionable distortions in complex images and their judgements are quite consistent. The same cannot be said about objective metrics, which, with a few exceptions,[4] require reference images. Secondly, objectionable and visible distortions are strongly correlated. The objectionable distortions can be approximated by visible distortions detected at lower sensitivity. In practice, it means that an objective visibility metric could potentially predict objectionable

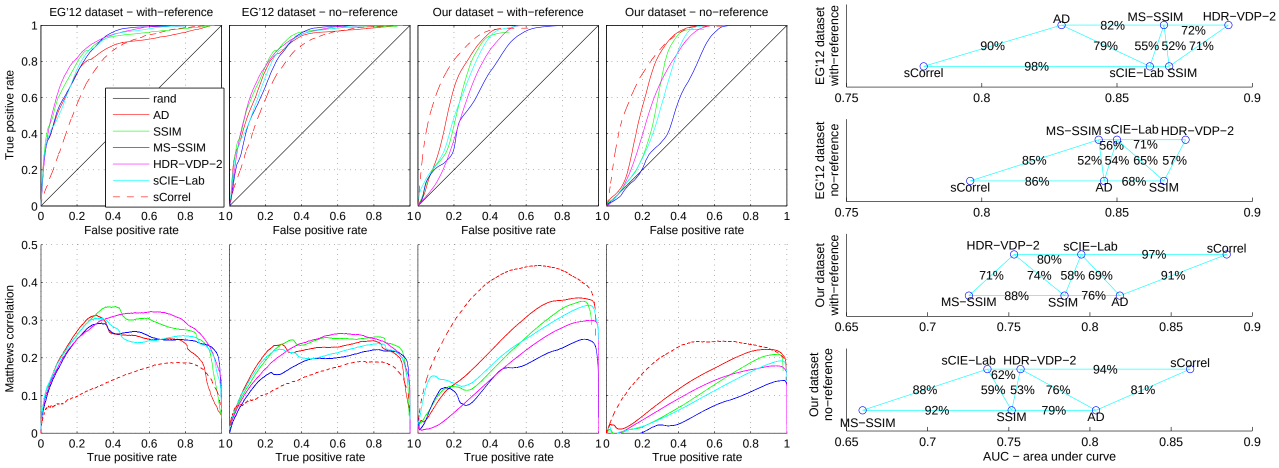


Figure 10: The performance of quality metrics shown as ROC plots [top-left], Matthews correlation [bottom-left] and ranked according to the area-under-curve (AUC) [right] (the higher the AUC, the better the classification into distorted and undistorted regions). The percentages indicate how frequently 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 1.3; sCIE-Lab — spatial CIELab; sCorrel — per-block Spearman's nonparametric correlation.

distortions when its sensitivity setting is set to a lower level.

The data set revealed weaknesses of both simple (PSNR, sCIE-Lab[31]) and advanced (SSIM, MS-SSIM,[19] HDR-VDP-2) quality metrics. The results for the two separate data sets we tested is shown in Figure 10. The most problematic shortcoming of the metrics was excessive sensitivity to brightness and contrast changes, which are common in graphics due to the bias of rendering methods (refer to Figure 11). The simple metrics failed to distinguish between imperceptible and well visible noise levels in complex scenes (refer to Figure 12). The multi-scale metrics revealed problems in localizing small-area and high-contrast distortions (refer to Figure 13). 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 14).

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 super-threshold distortions. Simple metrics, such as PSNR, can localize distortions well, but they fail to account for masking effects. We did not find evidence in our data set that any of the metrics, including PSNR, is significantly better than any other metric.

Our data set is available at: `http://www.mpi-inf.mpg.de/resources/hdr/iqm-evaluation/`. We believe that it has further potential in improving existing quality metrics, but also in analyzing the saliency of rendering distortions, investigating visual equivalence given our with- and no-reference data, and many other image quality-oriented applications.

3. PERSPECTIVE ON MODELING QUALITY IN GRAPHICS

In the future we can expect to find more work on visual metrics intended for graphics applications. This will include the metrics which account for other dimensions of the physical space of stimuli, such as color, depth, reflection properties, or light field. We can also expect to see metrics comparing geometrical representations and abstract properties rather than the resulting images.[32] As the metrics evolve, invariance to the changes that are not objectionable and have little impact on quality will become more important. The derivation of these metrics will become more data driven, where large data sets will be collected to calibrate, test and compare metrics. However, given enormous space of possible images and distortions, collecting such data sets will require a major effort.

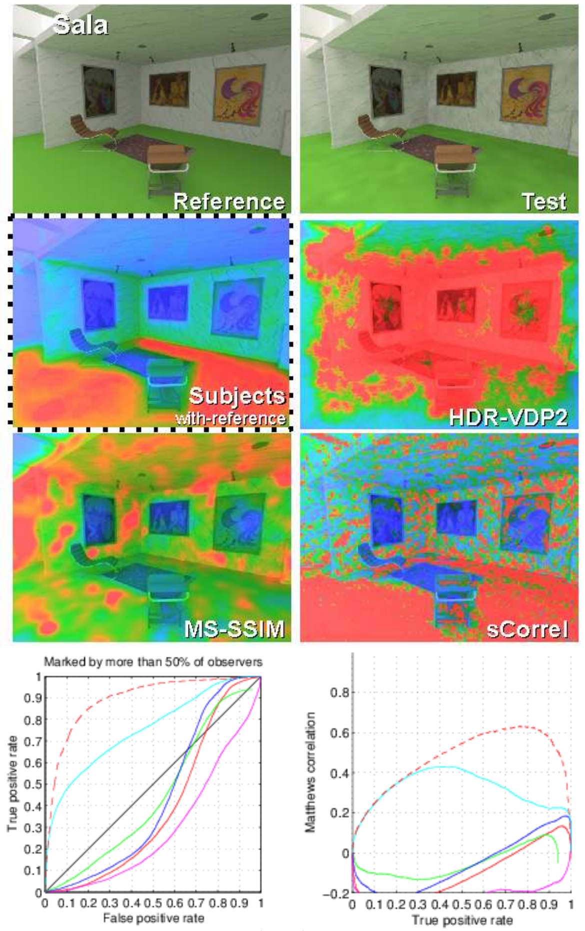


Figure 11: Scene *sala* (top), distortion maps for selected metrics (2^{nd} and 3^{rd} rows), ROC and correlation plots (bottom). Most metrics are sensitive to brightness changes, which often remain unnoticed by observers. *sCorrel* (block-wise Spearson correlation) is the only metric robust to these artifacts. Refer to the legend in Figure 10 to check which lines correspond to which metrics in the plots.

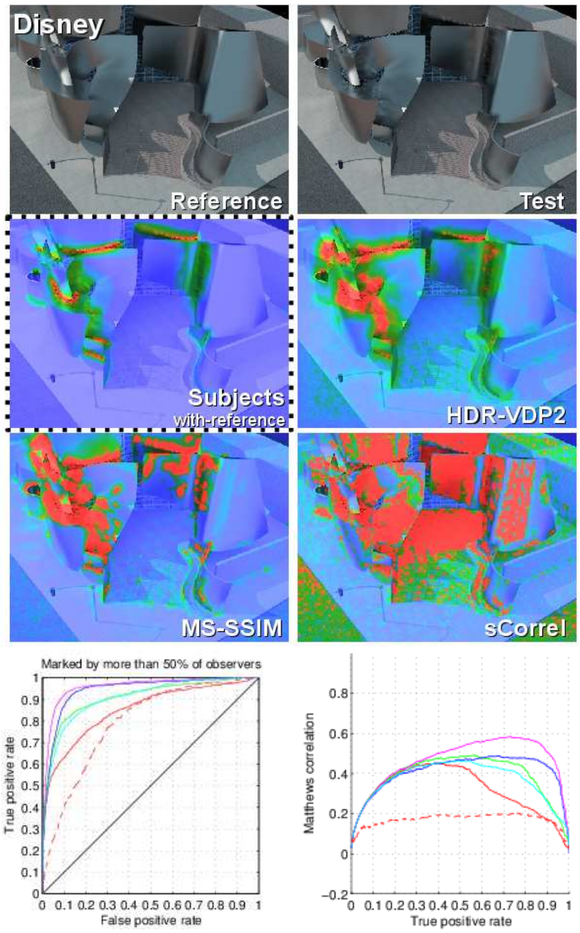


Figure 12: Scene *disney*: simple metrics, such as sCorrel and AD, fail to distinguish between visible and invisible amount of noise resulting in worse performance.

But even if such data can be collected, modeling the effects will pose a significant challenge. There are two main approaches to such modeling: 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 HDR-VDP-2 reviewed in Section 1.3 is an example of a white-box approach, while the data-driven metrics for non-reference quality prediction,[4] or color palette selection[23] 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 govern 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.

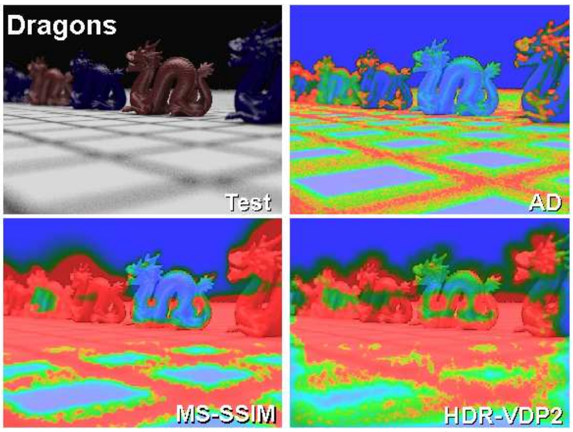


Figure 13: *Dragons* scene contains artifacts on the dragon figures but not in the black background. Multi-scale IQMs, such as MS-SSIM and HDR-VDP-2, mark much larger regions due to the differences detected at lower spatial frequencies. Pixel-based AD (absolute differences) can better localize distortions in this case.

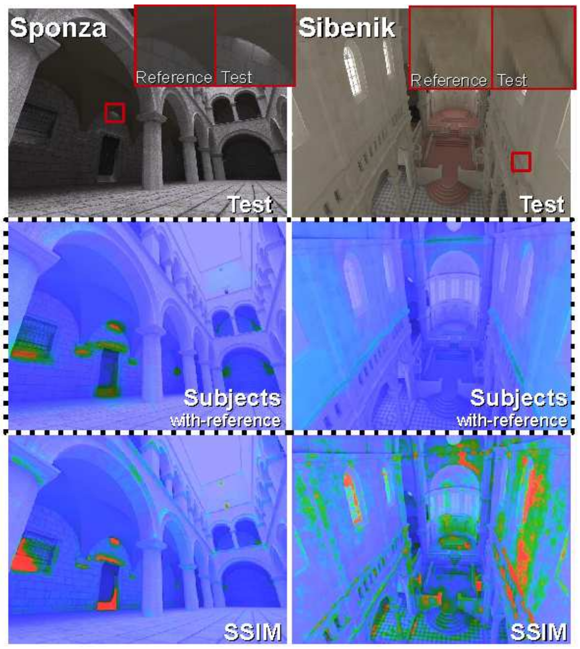


Figure 14: Photon leaking and VPL clamping artifacts in scenes *sponza* and *sibenik* result in either brightening or darkening of corners. Darkening is subjectively acceptable, whereas brightening leads to objectionable artifacts.

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. For example, many metrics make a simplifying assumption that the contrast perception can be modelled by the contrast sensitivity function (CSF) alone. However, the CSF has almost no impact on the super-threshold contrast, which dominates in complex images. Therefore, a CSF used alone in a visual model can actually worsen rather than improve metric predictions. 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 (refer to *aesthetics and naturalness* paragraph in Section 1.1), 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.

3.1 Conclusions

We found that the pairwise comparison methods, which reduce the number of comparison using efficient sorting algorithms, are likely to be the most time-efficient methods for subjective quality assessment. This is true given that the purpose of the experiment is to order the tested conditions rather than estimate interval-scale quality values; and that the criterion is reducing the time observers spent completing the experiment while maintaining the same level of test sensitivity.

We expect that device-independent imaging, offered by HDR color spaces will become more dominant in the near future. This will bring the need for quality metrics, which could operate on device-independent images. Such quality metrics will need to account for the fact that the perception of colors and contrast changes significantly across visible luminance range. In this paper we reviewed two metrics addressing this problem: the perceptually uniform luminance encoding (Section 1.1) and HDR-VDP-2 (Section 1.3).

Finally, we tested popular quality metrics against the data set consisting of spatially localized computer graphics distortions. We did not find evidence that any of the popular image quality metrics is significantly and universally better than any other metric, including PSNR. This could be the limitation of our data set of 27 images, which is unlikely to cover all types of images and distortions and provide sufficient statistical evidence. But more likely this is due to the fact that each metric fails to produce satisfactory predictions in at least a few cases.

ACKNOWLEDGMENTS

This work was partly supported by the EPSRC research grant EP/I006575/1.

REFERENCES

[1] Mantiuk, R. K., Tomaszewska, A., and Mantiuk, R., "Comparison of four subjective methods for image quality assessment," *Computer Graphics Forum* **31**(8), 2478–2491 (2012).

[2] Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W., "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph (Proc. SIGGRAPH)* **30**, 40 (July 2011).

[3] Cadík, M., Herzog, R., Mantiuk, R. K., Myszkowski, K., and Seidel, H.-P., "New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts," *Transactions on Graphics* **31**(6), 147 (2012).

[4] Herzog, R., Čadík, M., Aydčin, T. O., Kim, K. I., Myszkowski, K., and Seidel, H.-P., "NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis," *Computer Graphics Forum* **31**, 545–554 (May 2012).

[5] Giesen, J., Schuberth, E., Simon, K., Zolliker, P., and Zweifel, O., "Image-Dependent Gamut Mapping as Optimization Problem," *IEEE Transactions on Image Processing* **16**, 2401–2410 (Oct. 2007).

[6] Mantiuk, R. and Seidel, H., "Modeling a generic tone-mapping operator," *Computer Graphics Forum (Proc. of Eurographics)* **27**(2), 699–708 (2008).

[7] Bolin, M. R. and Meyer, G. W., "A perceptually based adaptive sampling algorithm," in [*Proc. of SIGGRAPH*], 299–309 (1998).

[8] Myszkowski, K., Rokita, P., and Tawara, T., "Perceptually-informed accelerated rendering of high quality walkthrough sequences," in [*Eurographics Workshop on Rendering*], **99**, 5–18 (1999).

[9] Ramasubramanian, M., Pattanaik, S. N., and 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*], 73–82, ACM Press, New York, New York, USA (1999).

[10] Ramanarayanan, G., Ferwerda, J., and Walter, B., "Visual equivalence: towards a new standard for image fidelity," *ACM Transactions on Graphics (TOG)* **26**(3), 76 (2007).

[11] Trentacoste, M., Mantiuk, R., Heidrich, W., and Dufrot, F., "Unsharp Masking, Countershading and Halos: Enhancements or Artifacts?," *Computer Graphics Forum* **31**, 555–564 (May 2012).

[12] Smith, K., Krawczyk, G., and Myszkowski, K., "Beyond tone mapping: Enhanced depiction of tone mapped HDR images," *Computer Graphics Forum* **25**(3), 427–438 (2006).

[13] Aydin, T. O., Mantiuk, R., Myszkowski, K., and Seidel, H.-P., "Dynamic range independent image quality assessment," *ACM Transactions on Graphics (Proc. of SIGGRAPH)* **27**(3), 69 (2008).

[14] Aydin, T. O., Čadík, M., Myszkowski, K., and Seidel, H.-P., "Video quality assessment for computer graphics applications," *ACM Transactions on Graphics* **29**, 1 (Dec. 2010).

[15] Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., and Seidel, H.-p., "A perceptual model for disparity," *ACM Transactions on Graphics* **30**, 1 (July 2011).

[16] Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., Seidel, H.-P., and Matusik, W., "A luminance-contrast-aware disparity model and applications," *ACM Transactions on Graphics* **31**, 1 (Nov. 2012).

[17] Yang, X., Zhang, L., Wong, T.-T., and Heng, P.-A., "Binocular tone mapping," *ACM Transactions on Graphics* **31**, 93:1–93:10 (July 2012).

[18] Aydn, T. O., Mantiuk, R., and Seidel, H.-P., "Extending quality metrics to full luminance range images," in [*Proceedings of SPIE*], 68060B–10, Spie (2008).

[19] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E., "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing* **13**, 600–612 (Apr. 2004).

[20] Mantiuk, R., Daly, S., Myszkowski, K., and Seidel, H., "Predicting visible differences in high dynamic range images: model and its calibration," in [*Human Vision and Electronic Imaging*], 204–214 (2005).

[21] Daly, S., "The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity," in [*Digital Images and Human Vision*], Watson, A. B., ed., 179–206, MIT Press (1993).

[22] Secord, A., Lu, J., Finkelstein, A., Singh, M., and Nealen, A., "Perceptual models of viewpoint preference," *ACM Transactions on Graphics* **30**, 1–12 (Oct. 2011).

[23] O'Donovan, P., Agarwala, A., and Hertzmann, A., "Color compatibility from large datasets," *ACM Transactions on Graphics* **30**, 1 (July 2011).

[24] Silverstein, D. and Farrell, J., "Efficient method for paired comparison," *Journal of Electronic Imaging* **10**, 394 (2001).

[25] Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F., "TID2008 - A database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics* **10**, 30–45 (2009).

[26] Blackwell, H., "Contrast thresholds of the human eye," *Journal of the Optical Society of America* **36**(11), 624–632 (1946).

[27] McCann, J. and Rizzi, A., "Veiling glare: the dynamic range limit of hdr images," in [*Proc. of HVEI XII*], **6492**, 649213–649213, International Society for Optics and Photonics (2007).

[28] Georgeson, M. A. and Sullivan, G. D., "Contrast constancy: deblurring in human vision by spatial frequency channels.," *J. Physiol.* **252**, 627–656 (Nov. 1975).

[29] Watson, A. and Ahumada Jr, A., "A standard model for foveal detection of spatial contrast," *Journal of Vision* **5**(9), 717–740 (2005).

[30] Sheikh, H., Sabir, M., and Bovik, A., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing* **15**(11), 3441–3452 (2006).

[31] Zhang, X. and Wandell, B. A., "A spatial extension of CIELAB for digital color-image reproduction," *Journal of the Society for Information Display* **5**(1), 61 (1997).

[32] Corsini, M., Larabi, M., Lavoué, G., Petík, O., Váša, L., and Wang, K., "Perceptual metrics for static and dynamic triangle meshes," in [*Eurographics 2012 - State of the Art Reports*], 135–157 (2012).