Video Quality Assessment for Computer Graphics Applications

Tunç Ozan Aydın* Martin Čadík* Karol Myszkowski* Hans-Peter Seidel*
MPI Informatik

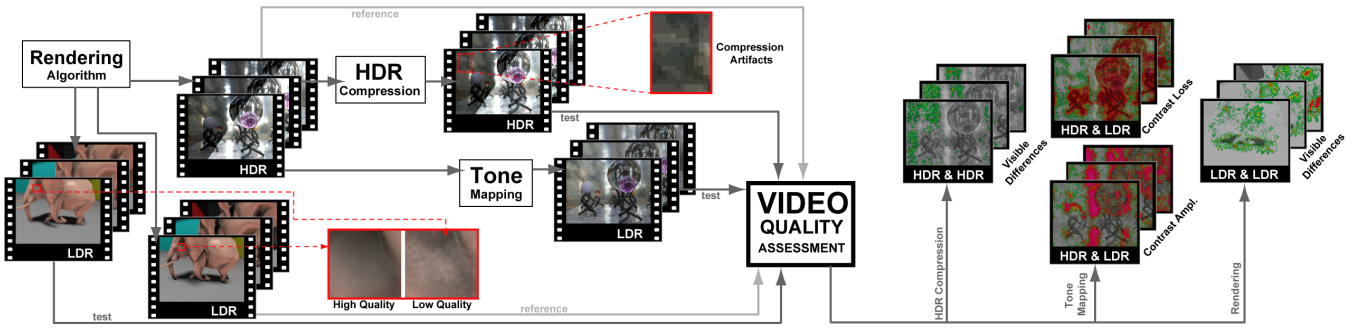

Figure 1: *The proposed metric predicts the perceived quality of natural as well as rendered video sequences with respect to a reference, even if the input videos have different dynamic ranges. Our work enables new applications including objective evaluation of video tone mapping and HDR compression.*

Abstract

Numerous current Computer Graphics methods produce video sequences as their outcome. The merit of these methods is often judged by assessing the quality of a set of results through lengthy user studies. We present a full-reference video quality metric[1] geared specifically towards the requirements of Computer Graphics applications as a faster computational alternative to subjective evaluation. Our metric can compare a video pair with arbitrary dynamic ranges, and comprises a human visual system model for a wide range of luminance levels, that predicts distortion visibility through models of luminance adaptation, spatiotemporal contrast sensitivity and visual masking. We present applications of the proposed metric to quality prediction of HDR video compression and temporal tone mapping, comparison of different rendering approaches and qualities, and assessing the impact of variable frame rate to perceived quality.

CR Categories: I.3.0 [Computer Graphics]: General; I.3.3 [Picture/Image Generation]: Display Algorithms—Viewing Algorithms

Keywords: video quality metrics, high dynamic range video, human visual perception, temporal artifacts, subjective video quality assessment

*e-mail: {tunc, mcadik, karol, hpseidel}@mpi-inf.mpg.de

[1] A web service that implements the metric described in this paper can be freely accessed at `http://drim.mpi-sb.mpg.de`.

1 Introduction

The contributions of newly proposed Computer Graphics techniques are usually demonstrated through images, and more often through videos, in which the merit of the technique is apparent. The performance of, for example a new rendering method, can be assessed by comparing sequences rendered on one hand using the proposed method, and on the other hand a more precise, but slower reference method. The point of this comparison could be to show that the proposed method produces results comparable to the reference method, but much more efficiently. A similar evaluation process is also common in other subfields such as High Dynamic Range (HDR) Imaging. Evaluation of tone mapping operators, as well as compression methods for HDR video both involve a comparison of, respectively the tone mapped and compressed video, with the HDR reference sequence. In fact, assessment of the fidelity of a video sequence to a reference is a task common to numerous Computer Graphics techniques.

Formal subjective methods of video quality evaluation such as [ITU-T 1999], where a Mean Opinion Score is computed by obtaining responses from multiple test subjects are often too laborious to be used on large sets of data. For the same reason the use of such methods in a feedback loop during development is not feasible; in fact most authors perform subjective evaluation only after the development of their algorithm is completed. Video Quality Metrics provide an objective means of comparing video sequences much faster than subjective methods by trading off accuracy of the prediction due to simplified modeling of visual perception. Simple metrics like PSNR, that rely solely on image pixel statistics fail to predict significant human visual system (HVS) properties like visual masking and contrast sensitivity. More sophisticated metrics [Winkler 2005; Seshadrinathan and Bovik 2010] on the other hand are not designed for HDR content. In the light of the recent trends towards HDR Imaging, the absence of HDR capable HVS models severely limits the use of these metrics in Computer Graphics context. Recently however, several *image* quality assessment metrics have been proposed, either designed specifically for HDR images [Mantiuk et al. 2005], or that can compare image pairs with arbitrary dynamic range [Aydın et al. 2008]. However, simply using image quality metrics to evaluate each frame of a video sequence fails to reflect the temporal aspects of Human Visual System's (HVS)

mechanisms, typically resulting in underestimating the visibility of temporal artifacts such as flickering (Sections 4, 5).

A video quality metric specifically designed for Computer Graphics applications by addressing the aforementioned issues, could be used as a practical diagnostic tool and a quick alternative to subjective evaluation. We propose a *dynamic range independent* video quality metric that can compare a video pair of arbitrarily different dynamic ranges. The metric comprises a temporal HVS model, that accounts for major effects like luminance adaptation, contrast sensitivity dependency to both spatial and temporal frequencies, and similarly visual masking computed in spatiotemporal visual channels (Section 3). Due to the absence of a visual attention model, the metric predictions are conservative in the sense that they correspond to the perception of an observer who scrutinizes the entire video sequence. The results in Section 4 show that our metric predicts distortion visibility more accurately than previous video quality metrics and state-of-the-art image quality assessment methods applied to each video frame separately. The predictions of the proposed metric are also validated through a subjective study (Section 5). We show that our metric enables new applications of evaluating HDR video tone mapping and compression methods. We also demonstrate the comparison of videos rendered with different methods and quality settings, and assessment of the impact of dropped frames to perceived quality (Section 6).

2 Background

In this section we summarize previous work on objective video quality assessment and the use of video quality measures in Computer Graphics applications, and give some background on the temporal HVS mechanisms related to our metric.

2.1 Video Quality Assessment

Video quality assessment metrics often draw ideas from the more developed image quality assessment field. It has been quickly observed that simple statistics like signal-to-noise ratio are not necessarily correlated with human vision, which motivated HVS-based image quality metrics. Commonly used image quality metrics focus on near-threshold detection [Daly 1993], supra-threshold discrimination [Lubin 1995], or functional differences [Ferwerda and Pellacini 2003]. The proposed video quality metric makes use of a near-threshold human visual system model to comply with the needs of computer graphics applications.

The focus of the early work on video metrics has been extending image quality assessment metrics with temporal models of visual perception, resulting from the fact that frame-by-frame application of image quality metrics is not sufficient. Van den Branden Lambrecht's Moving Picture Quality Metric (MPQM) [1996] utilizes a spatial decomposition in frequency domain using a filter bank of oriented Gabor filters, each with one octave bandwidth. Additionally two temporal channels, one low-pass (sustained) and another band-pass (transient) are computed to model visual masking. The output of their metric is a numerical quality index between $1-5$, similar to the Mean Opinion Score obtained through subjective studies. In a more efficient version of MPQM, the Gabor filter bank is replaced by the Steerable Pyramid [Lindh and van den Branden Lambrecht 1996]. In later work targeted specifically to assess the quality of MPEG-2 compressed videos [van den Branden Lambrecht et al. 1999], they address the space-time nonseparability of contrast sensitivity through the use of a spatiotemporal model. Another metric based on Steerable Pyramid decomposition aimed towards low bit-rate videos with severe artifacts is proposed by Masry and Hemani [2004], where they use finite impulse response filters for temporal decomposition.

Similarly, Watson et al. [2001] published an efficient Digital Video Quality metric (DVQ) based on the Discrete Cosine Transform. The DVQ models early HVS processing including temporal filtering and simple dynamics of light adaptation and contrast masking. Later they propose a relatively simple Standard Spatial Observer (SSO) based method [Watson and Malo 2002], which, on the Video Quality Experts Group data set, is shown to make as accurate predictions as more complex metrics. Winkler [1999; 2005] proposed a perceptual distortion metric (PDM) where he introduced a custom multiscale isotropic local contrast measure, that is later normalized by a contrast gain function that accounts for spatiotemporal contrast sensitivity and visual masking.

Seshadrinathan and Bovik [2007] proposed an extension to the Complex Wavelet Structural Similarity Index (CW-SSIM [Wang and Simoncelli 2005; Sampat et al. 2009]) for images to account for motion in video sequences. The technique (called V-SSIM) incorporates motion modeling using *optical flow* and relies on a decomposition through 3D Gabor filter banks in frequency domain. V-SSIM is therefore able to account for motion artifacts due to quantization of motion vectors and motion compensation mismatches. Recently, the authors published the MOVIE index in a follow-up work [Seshadrinathan and Bovik 2010], which outputs two separate video quality streams for every 16^{th} frame of the assessed video: *spatial* (closely related to the structure term of SSIM) and *temporal* (assessment of the motion quality based on optical flow fields). In Section 4 we compare our work with the MOVIE index and Winkler's PDM, along with a frame-by-frame evaluation by image quality metrics HDRVDP [Mantiuk et al. 2005] and the dynamic range independent metric [Aydın et al. 2008] (henceforth referred as DRIVDP).

2.2 Applications in Computer Graphics

The image quality evaluation with the use of HVS models has been an important topic in realistic image synthesis, particularly for static images [Rushmeier et al. 1995; Bolin and Meyer 1998]. More recently spatiotemporal models of visual perception have been considered for reducing the rendering time of animation sequences by exploiting limitations of the HVS. Myszkowski et al. [2000] proposed the use of an Animation Quality Metric (AQM), which utilizes image flow between a pair of subsequent frames to derive the retinal velocity, which is an input parameter for the spatiovelocity contrast sensitivity function (SVCSF) [Daly 1998]. Yee et al. [2001] further extended this work by using a computational model of visual attention to predict which image regions are more likely to be consciously attended by the observer, resulting in even more precise retinal velocity estimation. Both those techniques lack explicit processing of intensities between subsequent images, which makes detection of temporal artifacts such as flickering impossible. Such temporal information has been implicitly accumulated by averaging photon density across frame sequences and then applying the AQM metric to the resulting animation frames [Myszkowski et al. 2001]. However, in this case only temporal noise due to the photon density can be estimated, while other temporal artifacts such as flickering of improperly sampled textures or edge aliasing cannot be detected.

Schwarz and Stamminger [2009] propose a quality metric, which is targeted specifically for detection of popping artifacts due to level-of-detail (LOD) changes between frames. They assume the knowledge of the point in time when the LOD is changed and compare whether for that frame the differences for current and previous LOD (the latter image must be specifically re-rendered) are visible taking into account the SVCSF [Daly 1998]. Since temporal processing over frames is ignored, the influence of the dynamically changing scene and camera on the LOD change cannot be modeled prop-

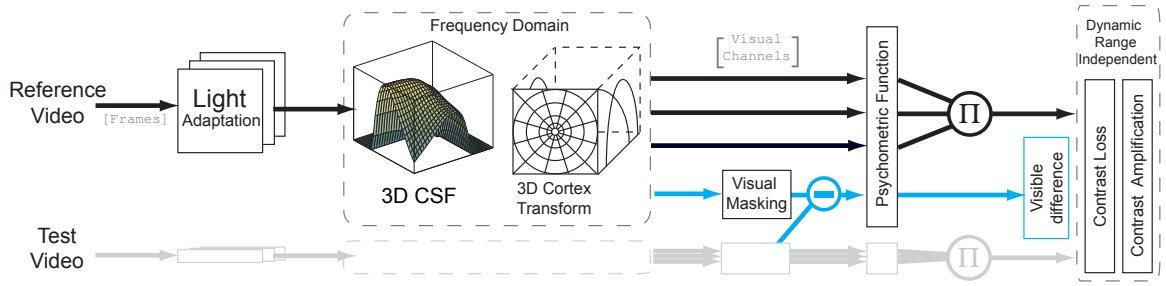

Figure 2: *The computational steps of our metric. Refer to text for details.*

erly. Clearly, an explicit 3D space-time contrast sensitivity function (CSF) processing over a number of subsequent frames is required to account for all possible temporal artifacts in a general setup, which is one of the main goals of our work.

2.3 Temporal Aspects of Human Visual System

Temporal Visual Channels
A significant area of interest of vision research is the Lateral Geniculate Nucleus (LGN), which is a portion of the brain inside the thalamus. It is estimated that 90% of monkey retinal ganglion cells send their axons to LGN layers, thus LGN is known as the primary processing center of visual information. In general, retinal ganglion cells can be divided into *midget* (smaller, majority of ganglion cells, sensitive to detail) and *parasol* (larger, faster output signals, sensitive to movement, only $\sim 10\%$) cells. LGN, in turn contains *parvocellular* (small cell bodies) and *magnocellular* (large cell bodies) layers. The axons of midget retinal ganglion cells terminate in the parvocellular layers, while the parasol cells terminate in magnocellular layers [Wandell 1995, p.124]. This structure suggests the existence of separate *parvocellular* and *magnocellular visual streams*.

Experiments have shown that the destruction of the cells in the parvocellular layers of a monkey's LGN resulted in deteriorated performance for a variety of tasks such as pattern detection and color discrimination. Destroying the cells in the magnocellular layers, however, did not affect the performance in the same tasks, but it was observed that the animal became less sensitive to rapidly flickering targets [Wandell 1995, p.126]. This leads to the conclusion that the magnocellular pathway is specialized to process high temporal frequency information [Watson 1986]. Meanwhile, some work has been done to find models that fit psychophysical measurements of the temporal sensitivity of human subjects. While models with many narrow band mechanisms, as well as three channels have been proposed in the past, it is now believed that there is just one low-pass, and one band-pass mechanism [Winkler 2005]. This theory is consistent with the biological structure of the LGN, moreover Friedericksen and Hess [1998] obtained a very good fit to large psychophysical data using only a *transient* and a *sustatined* mechanism.

Practical Implications
Although the parvo– and magnocellular pathways carry different types of information to the brain, the receptive fields of neurons in the parvocellular pathway are not space-time separable [Wandell 1995, p.143]. No clear anatomical separation between spatial and temporal frequencies supports the psychophysical finding that the contrast sensitivity is not separable along time and spatial dimensions. That leads to the **space-time nonseparability of the Contrast Sensitivity Function**. Thus, spatial CSFs measured for static stimuli cannot be extended linearly to account for the effect of temporal frequency to sensitivity. Another direct consequence of separate pathways for high and low temporal frequency contrast is the **spatiotemporal locality of inter-channel visual masking**. This suggests the use of 3D filter banks that span both spatial and temporal dimensions. Faithful modeling of temporal aspects of the HVS is vital in Computer Graphics applications, where flickering is an important source of visual artifacts. In Section 3 we describe how the proposed metric addresses these issues.

3 Video Quality Assessment

The recent proliferation of High Dynamic Range Imaging dictates that the HVS model employed in a video quality metric for Computer Graphics applications should be designed for all visible luminance levels. This requirement limits the use of earlier video quality metrics designed towards detecting compression artifacts in low dynamic range (LDR) videos. Moreover, applications such as tone mapping and compression of HDR video sequences require detecting structural distortions where the reference video is HDR and the test video is LDR. Consequently, in this work we use an HDR capable model that accounts for both major spatial and temporal aspects of the visual system, and employ the dynamic range independent distortion measures *contrast loss* and *amplification* introduced in DRIVDP in addition to simply computing the *visible differences* between reference and test videos. The HDR capability is a result of the light adaptation computation through the JND space transformation and the 3D contrast sensitivity function, both explained in more detail later in this section. In Computer Graphics applications the main concern is often the existence of visible artifacts, rather than the magnitude of visibility, since methods that produce clearly visible artifacts are often not useful in practice. Consequently the HVS model we use trades off supra-threshold precision for accuracy near the detection threshold.

The computational steps of our metric are summarized in Figure 2. The input is a pair of videos V_{ref} and V_{tst} with arbitrary dynamic ranges, both of which should contain calibrated luminance values. The luma values of LDR videos should be inverse gamma corrected and converted to display luminance (In all examples we assumed a display device with the luminance range $0.1 - 100\ cd/m^2$ and gamma 2.2). The HVS model is then applied separately to both videos to obtain the normalized multichannel local contrast at each visual channel, where the first step is to model the nonlinear response of the photoreceptors to luminance, namely **Light adaptation**. In our metric we apply the nonlinearity described in [Mantiuk et al. 2005], which maps the video luminance to linear Just Noticeable Differences (JND) values, such that the addition or subtraction of the unit value results in a just perceivable change of relative contrast[2].

[2] All externally referred derivations and formulas in the rest of the paper are recollected in supplementary material for easy reference.

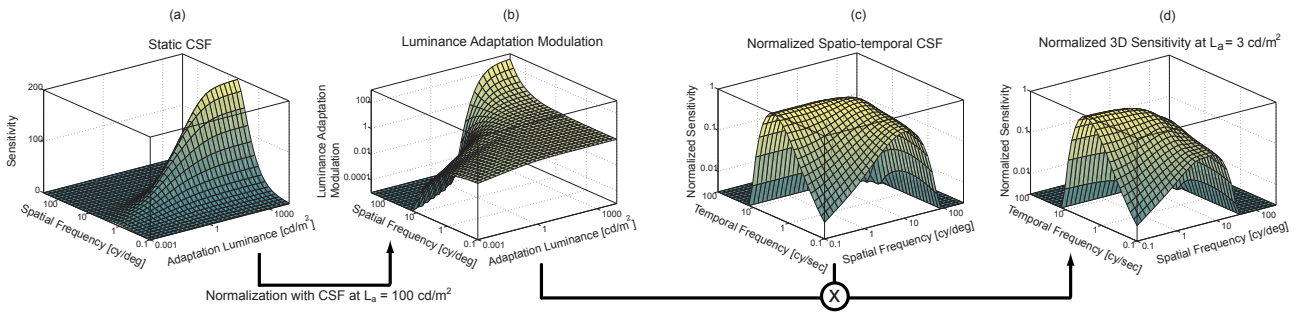

Figure 3: *Computation of the CSF^{3D}. The static $CSF^S(\rho, L_a)$ (a) is divided to $CSF^S(\rho, L_a = 100cd/m^2)$ to obtain scaling coefficients (b) that account for luminance adaptation in CSF^{3D}. The specific adaptation level is chosen to reflect the conditions where the spatiotemporal CSF^T was measured (c). The scaling coefficients are computed for the current L_a ($3\ cd/m^2$ in this case), and multiplied with the normalized CSF^T to obtain the CSF^{3D} that accounts for spatial and temporal frequencies, as well luminance adaptation (d).*

Contrast sensitivity is a function of spatial frequency ρ and temporal frequency ω of a contrast patch, as well as the current adaptation luminance of the observer L_a. The spatiotemporal CSF^T plotted in Figure 3c shows the human contrast sensitivity for variations of ρ and ω at a fixed adaptation luminance. At a retinal velocity v of $0.15\ deg/sec$, the CSF^T is close to the static CSF^S [Daly 1993] (Figure 3a) at the same adaptation level (the relation between spatio-temporal frequency and retinal velocity is $\omega = v\rho$ assuming the retina is stable). This particular retinal velocity corresponds to the lower limit of natural drift movements of the eye which are present even if the eye is intentionally fixating in a single position [Daly 1998]. In the absence of eye tracking data we assume that the observer's gaze is fixed, but also the drift movement is present. Accordingly, a minimum retinal velocity is set as follows:

$$CSF^T(\rho, \omega) = CSF^T(\rho,\ max(v, 0.15) \cdot \rho). \qquad (1)$$

In addition to the drift movement, one could consider integrating a visual attention model-based smooth pursuit eye motion (SPEM) estimate [Yee et al. 2001] (which may not always be precise), or actual eye tracking data to our metric, at the cost of introducing user input and thus loosing objectivity of the approach.

On the other hand, the shape of the CSF depends strongly on adaptation luminance especially for scotopic and mesopic vision, and remains approximately constant over $1000\ cd/m^2$. Consequently, using a spatiotemporal CSF at a fixed adaptation luminance results in erroneous predictions of sensitivity at the lower luminance levels that can be encoded in HDR images. Thus, we derive a "3D" CSF (Figure 3d) by first computing a *Luminance Modulation Factor* (Figure 3b) as the ratio of CSF^S at the observer's current adaptation luminance (L_a) with the CSF^S at $L_a = 100\ cd/m^2$, which is the adaptation level at which the CSF^T is calibrated to the spatiotemporal sensitivity of the HVS. This factor is then multiplied with the normalized spatiotemporal CSF ($nCSF^T$), and finally the resulting CSF^{3D} accounts for ρ, ω and L_a:

$$CSF^{3D}(\rho, \omega, L_a) = \frac{CSF^S(\rho, L_a)}{CSF^S(\rho, 100)} nCSF^T(\rho, \omega). \qquad (2)$$

Ideally the CSF^{3D} should be derived from psychophysical measurements in all three dimensions, since current findings suggest that the actual contrast sensitivity of the HVS is linearly separable in neither of its dimensions. In the absence of such measurements, we found that estimating luminance adaptation using a scaling factor is better than the alternatives that involve an approximation by linear separation of spatial and temporal frequencies (as discussed earlier in Section 2.3). The effect of luminance adaptation to spatiotemporal contrast sensitivity can approximately be modeled by a multiplier (Figure 3b) except for very low temporal frequencies [Wandell 1995, p.233].

The perceptually scaled luminance contrast is then decomposed into *visual channels*, each sensitive to different temporal and spatial frequencies and orientations. For this purpose we extend the **Cortex Transform** [Watson 1987] that comprises 6 spatial frequency channels each further divided into 6 orientations (except the base band), by adding a sustained (low temporal frequency) and a transient (high temporal frequency) channel in the temporal dimension (total 62 channels). The time (t given in seconds) dependent impulse responses of the sustained and transient channels, plotted in Figure 4-left, are given as Equation 3 and its second derivative, respectively [Winkler 2005]:

$$f(t) = e^{-\frac{\ln(t/0.160)}{0.2}}. \qquad (3)$$

The corresponding frequency domain filters are computed by applying the Fourier transform to both impulse responses and are shown in Figure 4-right.

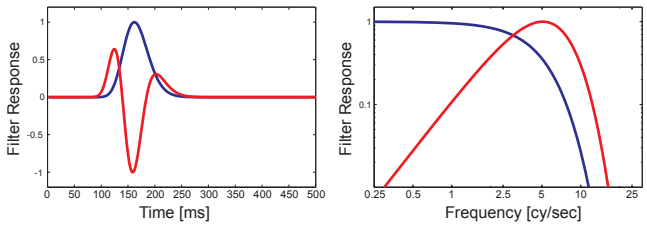

Figure 4: *Impulse (left) and frequency (right) responses of the transient (red) and sustained (blue) temporal channels. The frequency responses comprise the extended 3D Cortex Transform's channels in temporal dimension.*

Combining all models discussed so far, the computation of visual channels from the calibrated input video V is performed as follows:

$$C^{k,l,m} = \mathscr{F}^{-1}\left\{V_{csf}\ cortex^{k,l} \times temporal^m\right\} \text{ and}$$
$$V_{csf} = \mathscr{F}\{jnd(V)\}\ CSF^{3D},$$

where the 3D Cortex Filter for channel $C^{k,l,m}$ is computed from the corresponding 2D cortex filter $cortex^{k,l}$ at spatial frequency level

k and orientation l, and the sustained and transient channel filters $temporal^m$. The function jnd denotes the light adaptation nonlinearity, and $\mathscr{F}$ is the Fourier Transform. The threshold elevation due to **visual masking** is computed using the following nonlinearity [Daly 1993]:

$$Te^{k,l,m} = \left[1 + \left(0.0153\left(392.498|C_{pu}^{k,l,m}|\right)^{slope}\right)^4\right]^{\frac{1}{4}}, \quad (4)$$

where $C_{pu}^{k,l,m}$ indicates the channel with *phase uncertainty* and the $slope$ is linearly interpolated between $0.7 - 1$ for visual channels from low to high spatial frequencies.

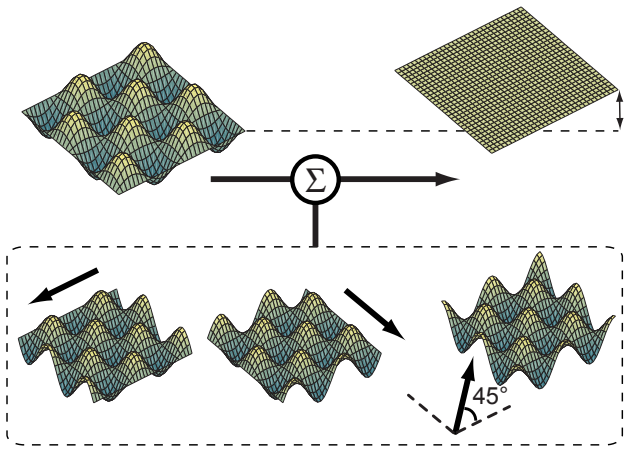

Figure 5: *Practical illustration of achieving phase uncertainty in 2D. The Hilbert transform should be applied in multiple orientations to obtain a phase independent signal.*

The dependency of the visual channels to signal phase contradicts with the observation that the phase sensitivity of the HVS is very limited. Phase uncertainty, while often not explicitly mentioned, is a crucial component of many quality assessment metrics. If one uses a decomposition consisting of spatially even filters, the filter responses would contain zero crossings at step edge locations. This contradicts with human perception which exhibits a strong response to step edges. Analogously, in the temporal dimension sudden changes in pixel intensity are perceived strongly. The effect of phase uncertainty on complex stimuli is often a reduced amount of detected distortions, due to the increased visual masking in step edge locations. A common way of removing phase dependency of a 1D signal is to use a *quadrature pair* of filters where one filter is obtained by shifting the other's phase by 90 degrees. Although the phase shift can be computed in 1D by means of Hilbert transform, the extension of the Hilbert transform to higher dimensions is not trivial (Figure 5). Our implementation of phase uncertainty is an extension of the quadrature cortex filters [Lukin 2009] to the temporal domain. The spatial phase-shift is computed using an oriented 2D Hilbert Transform:

$$h^S(\rho_x, \rho_y) = i\ sgn(p\ \rho_x + q\ \rho_y), \quad (5)$$

where i is the imaginary unit, and the line given by the equation $p\ \rho_x + q\ \rho_y = 0$ specifies the "direction" of the transform. Parameters p and q are selected such that the direction of the Hilbert Transform coincides with the spatial orientation of the cortex channel. In the temporal dimension the phase shift can be achieved using a 1D Hilbert Transform:

$$h^T(\omega) = i\ sgn(\omega). \quad (6)$$

The quadrature responses of spatiotemporal visual channels are then computed as follows:

$$H^{S|T}\{C^{k,l,m}\} = \mathscr{F}^{-1}\{h^{S|T}\ \mathscr{F}\{C^{k,l,m}\}\}. \quad (7)$$

The phase independent channel $C_{pu}^{k,l,m}$ used in the threshold elevation formula is computed by summing up the original signal with all phase shifted responses in spatial and temporal dimensions as illustrated in Figure 6.

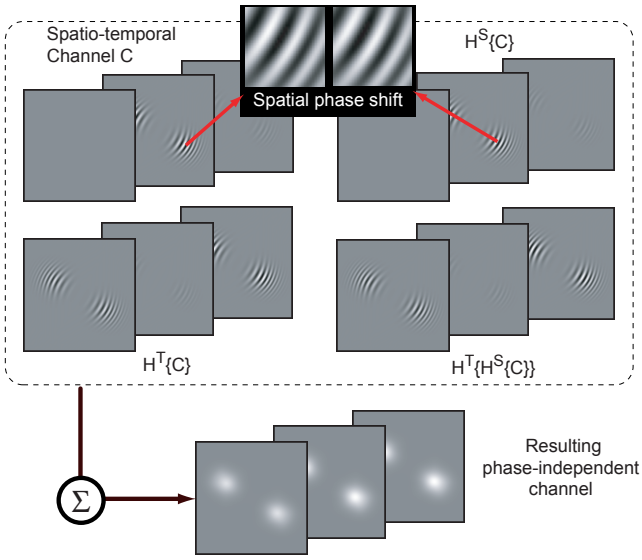

Figure 6: *3D phase uncertainty on a frequency plate image modulated in temporal domain using a sinusoid function. The spatiotemporal channel C obtained by 3D Cortex Transform is used to compute $H^S\{C\}$, $H^T\{C\}$ and $H^T\{H^S\{C\}\}$, the phase shifted response in spatial, temporal and both dimensions, respectively. The combination of all four responses yields a spatiotemporaly phase independent response constant along the entire sequence.*

The detection probability of the normalized contrast response C at each visual channel is computed using the following **psychometric function**, separately for the reference and test images:

$$P(C) = 1 - \exp(-|C|^3). \quad (8)$$

The psychometric function relates the normalized contrast to detection probability. Using this function, we compute the detection probabilities of the following three types of distortions:

- **Visible Difference** $\left(P_{\Delta}^{k,l,m} = P(\frac{C_{tst}^{k,l,m}}{Te_{tst}^{k,l,m}} - \frac{C_{ref}^{k,l,m}}{Te_{ref}^{k,l,m}})\right)$
- **Contrast Loss** $\left(P_{\searrow}^{k,l,m} = P(C_{ref}^{k,l,m})(1 - P(C_{tst}^{k,l,m}))\right)$
- **Contrast Amplification** $\left(P_{\nearrow}^{k,l,m} = P(C_{tst}^{k,l,m})(1 - P(C_{ref}^{k,l,m}))\right)$

The visible differences between video sequences convey more information than the other two types of distortions, but especially if the input video pair has different dynamic ranges, the probability map is quickly saturated by the contrast difference that is not necessarily perceived as a distortion. In this case contrast loss and amplification are useful which predict the probability of a detail visible in the reference becoming invisible in the test video, and vice versa.

While additionally contrast reversal proposed in DRIVDP can be easily computed within this framework, we found that this type of distortion did not convey further information in the examples we considered, and thus excluded from the metric output. Detection probabilities of each type of distortions are then combined using a standard probability summation function:

$$\hat{P}_{\Delta|\searrow|\nearrow} = 1 - \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{m=1}^{M} \left(1 - P_{\Delta|\searrow|\nearrow}^{k,l,m}\right). \quad (9)$$

The resulting three *distortion maps* $\hat{P}$ are visualized separately using an in-context distortion map approach where detection probabilities are shown in color over a low contrast grayscale version of the test video. We also found that an overall summary of the distortion information conveyed through a 3D visualization is useful in certain applications (Section 6.4).

4 Results

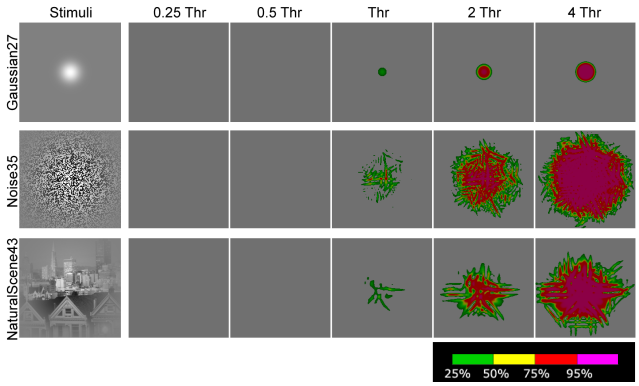

Figure 7: *Predicted visible differences between selected stimuli from the Modelfest data set and the background luminance, where the stimuli is scaled at $\frac{1}{4}, \frac{1}{2}, 1, 2$ and 4 times the threshold contrast (The same color coding is used throughout the paper for visualizing distortion detection probabilities, unless noted otherwise).*

In this section we compare the predictions of our metric with the outcomes of the recent video quality metrics PDM [Winkler 2005] and the MOVIE index [Seshadrinathan and Bovik 2010]. Although not intended for videos, we also considered two recent HDR capable image quality metrics HDRVDP [Mantiuk et al. 2005] and DRIVDP [Aydın et al. 2008], with which we evaluated each video frame separately. To ensure that our metric is calibrated to psychophysically measured detection thresholds, we computed the visible differences of the Modelfest data set at five different contrast levels with the background luminance. The video for a stimulus is generated by repeating it in all frames. As expected, the majority of the stimuli produced no response below the threshold, and a response with increasing magnitude for near– and above threshold. Figure 7 shows the outcome for selected stimuli relevant to our applications: a low and a high frequency noise, and a complex image. The worst results were obtained for "GaborPatch9" and "Gaussian26" for which our metric was too insensitive[3].

The test video for this section is generated using an HDR image, to which we added spatiotemporal random noise filtered with a Gaussian to roughly mimic the artifacts that appear in rendered videos in the absence of temporal coherency. The magnitude of the noise has been modulated with the luminance levels of the relatively dark image that depicts a sunset. The reference video is generated similarly by repeating the same HDR image in all frames. The frames in Figure 8, tone mapped using Pattanaik's operator [2000], depict the approximate appearance of the scene.

Figure 8: *Approximate perception of the reference and test scenes*

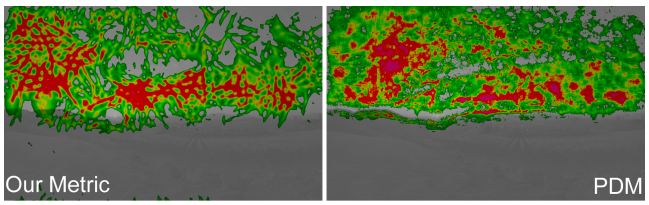

Figure 9: *Metric comparison for LDR test and reference videos*

First, we compare the distortion visibility prediction of our metric with PDM and MOVIE index on this tone mapped LDR image pair. Due to the random nature of the distortion, the frames of the distortion maps in this section are very similar, and thus we arbitrarily choose a single representative frame[4]. In this case the outcome of our metric and the PDM are similar (Figure 9).

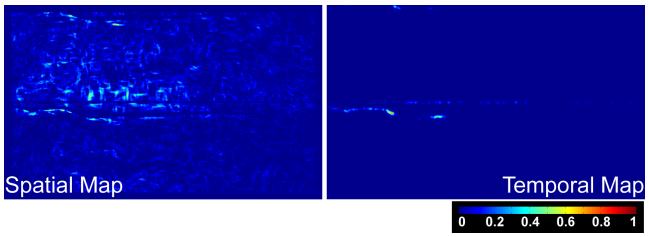

Figure 10: *MOVIE index for LDR videos. Note the different color coding*

The output of the MOVIE index on the other hand are a series of spatial and a temporal distortion maps that are computed at every 16^{th} frame. In Figure 10 we show the spatial distortion map at the 3^{rd} scale along with the temporal distortion map. While the output format of the MOVIE index is not directly comparable with other metrics discussed in this section, one can see that the spatial map of structural distortions (Figure 10-left) closely correlates to the distortions in the video sequence. However, due to the lack of a mechanism to estimate threshold contrast, distortions are detected even at the darker bottom half of the video.

Next, we test the metrics on the HDR test and reference videos. Note that the HDR format is capable of encoding the actual scene luminance unlike display-referred LDR videos in the previous case. The MOVIE index is excluded from the remaining comparisons

[3] Refer to supplementary material for responses to all Modelfest stimuli.

[4] All original video sequences and corresponding distortion maps are presented in the supplementary video.

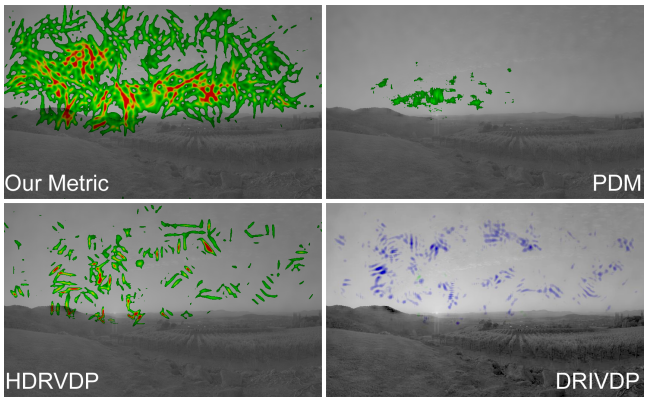

Figure 11: *Metric comparison for HDR test and reference videos. The contrast amplification in DRIVDP is color coded with blue.*

since its extention to HDR is not trivial. The difference in predictions of our metric and PDM in this case is because the latter does not model luminance adaptation. Consequently distortion visibility is underestimated due to artificially high thresholds in this low luminance scene (Figure 11). The visible difference and contrast amplification predicted by frame-by-frame evaluation of HDRVDP and DRIVDP are also noticeably lower than ours due to the absence of a temporal model that accounts for the higher sensitivity to flickering distortions compared to static distortions.

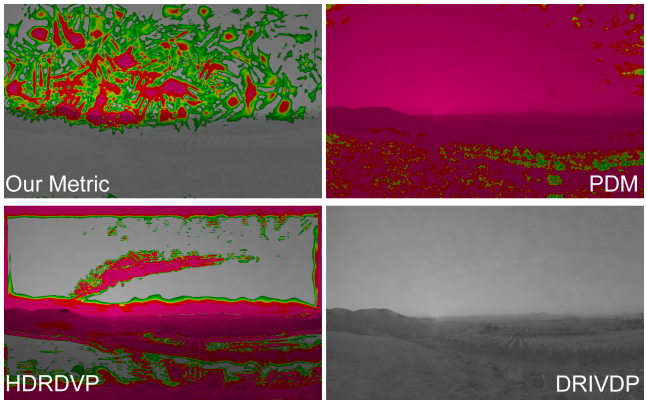

Figure 12: *Metric comparison for HDR reference and LDR test videos*

An even more striking difference can be observed in the final setup where the distorted video tone mapped with Pattanaik's operator is compared with the reference HDR video (Figure 12). Here, both PDM and and HDRVDP's distortion maps are dominated by the contrast difference due to the different dynamic ranges of the input video pair. This is especially evident in HDRVDP's prediction where the spatiotemporal distortion appears to be completely ignored. Moreover, DRIVDP predicts no visible detail amplification at all, since it does not detect the distortion and is also not affected by the different dynamic ranges of the input videos. The contrast amplification predicted by our metric on the other hand correctly identifies distortions where they are visible, and similar to DRIVDP also ignores the changes due to dynamic range difference. Note also that the predictions of our metric in all three scenarios are fairly consistent.

5 Validation

We performed a subjective study to validate the prediction performance of the metric[5]. The metric's capability of working on video pairs with different dynamic ranges, as well as the outcome in the form of distortion maps containing spatial information, demanded the creation of a new data set, since current public video quality databases are limited to LDR videos, and the measured subjective data is a single number indicating overall quality without any information on spatial distribution of visible distortions. To that end, a test set of 9 reference-test video pairs (1 LDR-LDR, 2 HDR-LDR, and 6 HDR-HDR) were generated by adding temporally and spatially varying artifacts (such as random noise, compression, tone mapping and luminance modulation) to 6 different HDR scenes. A BrightSide DR37-P HDR display was employed to properly display the scene luminance of both HDR and LDR videos. The participants of the study were 16 subjects between ages 23 and 50, all with near perfect or corrected vision. They were shown all video pairs side by side on the HDR display, and were asked to mark the visible differences (detail loss and amplification for HDR-LDR stimuli) on a 16×16 grid displayed over the video using a graphical user interface (Figure 13).

Figure 13: *The graphical user interface displays the test video (left) side-by-side with the corresponding reference video. The subjects mark regions where they notice visible differences on a 16×16 grid (right). Both video frames are tone mapped, and the distortions in the left frame are exaggarated for illustration purposes.*

The marked regions for each trial were stored as distortion maps, which were then averaged over all subjects to find the mean subjective response. Next, the metric prediction for the corresponding stimulus was computed, averaged over all frames, and downsampled to the same resolution as the mean subjective response. For each video pair, we computed the 2D correlation between the mean subjective response and the metric prediction. The correlations varied from 0.733 to 0.883, averaging to 0.809. The high correlation between the metric predictions and subjective responses over a diverse test set including HDR and LDR stimuli with distortions of various type and magnitude indicate that the proposed metric provides a reliable estimate of the video quality as a function of spatial location. For comparison, we also evaluated the test set with PDM, HDRVDP and DRIVDP (Figure 14). For almost all stimuli our metric's predictions were more accurate with respect to the subjective data, and the average correlations over all stimuli were found as 0.257 for PDM, 0.528 for HDRVDP, and 0.563 for DRIVDP.

[5]Refer to the supplementary material for a detailed discussion of the experiment.

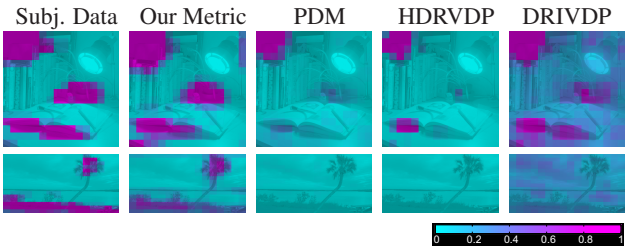

Figure 14: *The comparison of the subjective data averaged over participants with the predictions of our metric, PDM, HDRVDP and DRIVDP for stimulus #2 and #4 in our test set (refer to the supplementary material for the complete set of results).*

6 Applications

The proposed method for objective quality assessment of a test video with respect to a reference without any constraints on the dynamic range provides a faster alternative to subjective evaluation of rendering methods, and also enables a computational comparison of HDR video compression and tone mapping techniques. We also show that our metric gives insight on the effect of dropped frames to overall quality.

6.1 HDR Video Compression

While HDR content is becoming more commonplace, since it offers higher fidelity compared to traditional media, it does so at the cost of significantly increased file sizes. This is often not a problem for images due to cheaply available storage. However, working with long, high resolution videos quickly becomes prohibitively expensive. Incidentally HDR video compression has become an active topic of research. Figure 15 shows that our metric can be used to detect compression artifacts in a video sequence compressed [Mantiuk et al. 2004] at various quality settings.

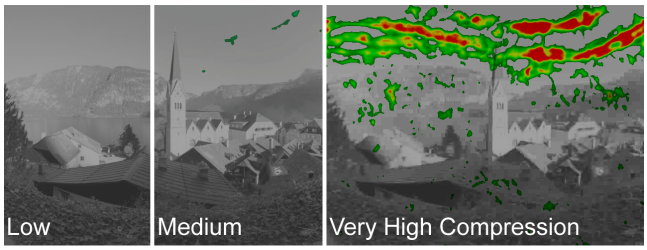

Figure 15: *Visible differences between frames from the HDR video and the corresponding compressed frames shown in three compression settings (Low – q=1, Medium – q=5, Very High – q=31). The banding artifacts become clearly visible under extreme compression. Near the foliage at the bottom, banding artifacts are present but not visible due to the low luminance*

6.2 Temporal Tone Mapping

HDR display technology is still early in its development, thus it is often necessary to reduce the dynamic range of the HDR content such that it can be viewed on current display hardware. While the goal of tone mapping is considered to be subjective, the fidelity of the tone mapped video to the reference HDR is often a good indicator of quality. In Figure 16 we show the results from selected frames of a tone mapped HDR sequence computed with global [Drago et al. 2003] and gradient based [Fattal et al. 2002] tone mapping methods.

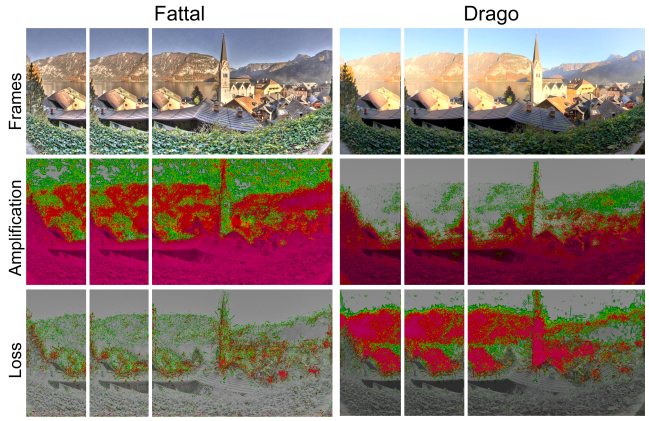

Figure 16: *Selected frames from the tone mapped HDR sequences and corresponding contrast amplification and loss maps. Each frame of the reference HDR video is tone mapped separately. Fattal's gradient based operator enhances perceived contrast notably, thus leading to highly detectable contrast amplification but little contrast loss. Drago's global operator on the other hand produces a more "flat" image by amplifying contrast near the dark foliage in the foreground and clipping brighter details near the horizon line.*

Another interesting practical problem involves both temporal tone mapping and compression. Consider a scenario where visual content is stored in a centralized media server in compressed HDR format. One may require to perform on-the-fly tone mapping to reduce the video's dynamic range to be suitable for the client machine's display device, which may range from an high-end LCD panel to a limited CRT monitor. An obvious consideration in this case is to make sure that tone mapping does not amplify previously invisible compression artifacts. In Figure 17 we show such an example where tone mapping adversely affects perceived quality of the compressed HDR video, which is correctly detected by our metric.

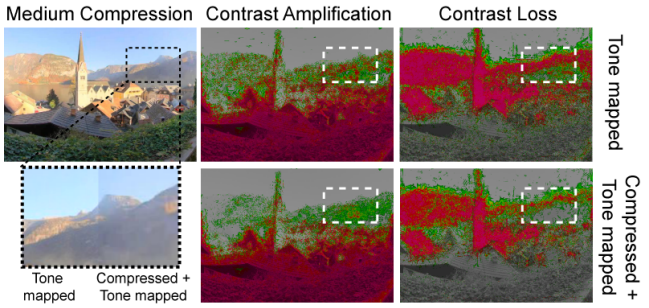

Figure 17: *Contrast amplification and loss predicted with respect to the reference HDR sequence for the compressed (at medium quality) and then tone mapped sequence using Drago's operator. Note the slightly increased contrast amplification and loss in the tone mapped version of the compressed HDR video. As shown in Figure 15, the artifacts generated in medium compression setting for this scene are mostly not detectable in the HDR video, but they become visible due to tone mapping applied later.*

6.3 Rendering

Our metric can be used to compare different rendering approaches. Figure 18 shows the visible differences of a dynamic scene walkthrough rendered with indirect lighting using reflective shadow maps [Dachsbacher and Stamminger 2005] with 1000 virtual point light (VPL) sources, with respect to the reference sequence obtained with the same amount of VPLs, however using a recent technique [Herzog et al. 2010] that utilizes spatio-temporal filtering. Due to this filtering, there are virtually no visible artifacts in the reference sequence, while the test technique produces visible flickering during the entire sequence.

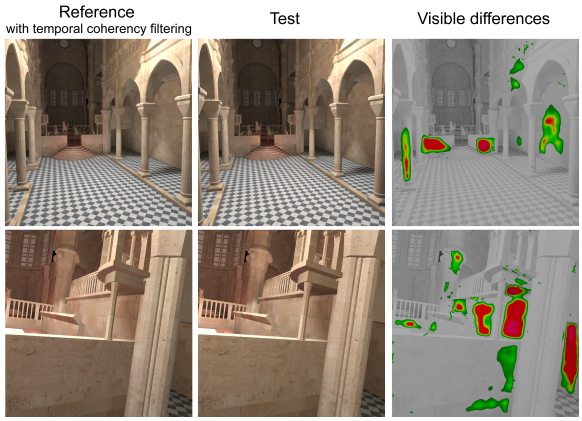

Figure 18: *Visible differences between rendering techniques. Even though the rendered frames are visually indistinguishable when viewed side-by-side, the test method produces significantly visible flickering artifacts, which is not the case for the reference method with temporal coherency filtering. Our metric also detects the non-uniform perception of these flickering artifacts, such as the perception of the artifacts on the ground masked by the moving checkerboard pattern (better visible in the supplementary video).*

To complement the previous scene with mostly temporal distortions, we show another example with artifacts of spatiotemporal nature (Figure 19). Here, the squences are rendered using an image-space horizon based ambient occlusion technique [Bavoil et al. 2008] augmented with the screen space directional occlusion (SSDO) [Ritschel et al. 2009] (48×32 and 12×10 polar samples on the hemisphere for the reference and test sequences, respectively) with directional light source sampled from an environment map (128 and 96 samples, respectively) and percentage closer filtering (PCF) shadow maps [Reeves et al. 1987] (64 and 16 samples, respectively). Visible differences are predicted mostly near the boundaries of the elephant's shadow.

6.4 Variable Frame Rate

Maintaining a high enough frame rate is desirable in applications like rendering and video streaming, but at the same time is not always possible due to hardware or bandwidth limitations. In this case, the visible differences between the low FPS video and the full FPS reference is a good measure for the loss in perceived quality due to low frame rate. Figure 20 shows that our metric can be used to predict the perceived distortions caused by dropped frames in a rendered walkthrough scene. The reference sequence was generated by Coherent Hierarchical Culling technique [Bittner et al. 2004] which never falls below 60 FPS for this scene. On the other hand, the performance of the traditional view frustum culling drops below 1 FPS at times. We also show an alternative 3D visualization

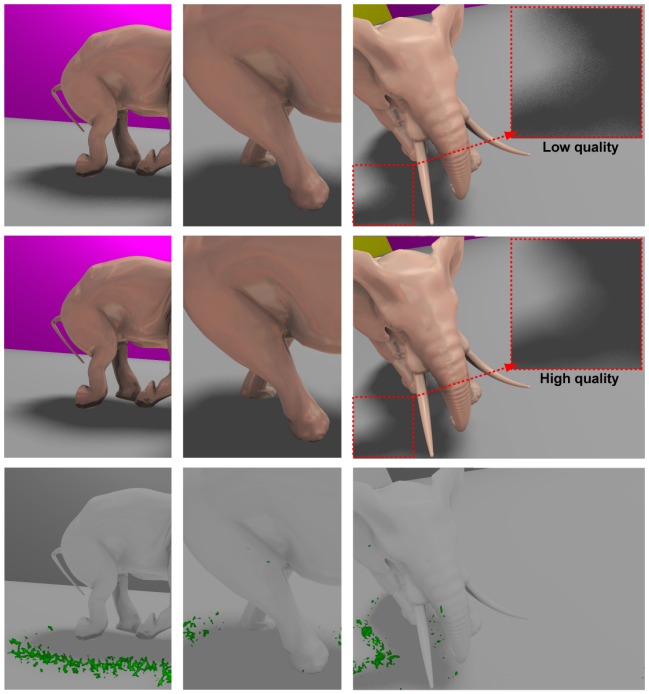

Figure 19: *Visible differences (bottom row) between the high (top row) and low quality (middle row) renderings are focused mostly near shadow boundaries.*

of this scene utilizing volume rendering that gives an overview of the distortion data (Figure 21). Note that the perception of frame freezes and drops has further aspects (e.g. judder) that are not accounted for by our method.

7 Discussion

The running time of the proposed metric depends highly on the resolution and length of the input videos, however in its current state is intended to work offline (~ 5 minutes for $512 \times 512 \times 64$ sequence). In our experience, the main bottleneck in performance is computing the 3D Fourier Transform of an 64 frames portion of the video, where that specific number is chosen because the sensitivity to temporal frequencies higher than 32 cy/sec is significantly low. This approach also requires that the portions of the video being processed should be kept in memory.

While our implementation runs in a standard workstation hardware without problems, another approach that trades off efficiency for prediction accuracy is to approximate the frequency domain Cortex Transform with the Steerable Pyramid decomposition performed in the spatial domain through polynomial approximations of the second derivative Gaussian filters [Freeman and Adelson 1991]. The filters that compute transient and sustained temporal channels can also be approximated by 9-tap filters corresponding to the impulse responses given in Figure 4 as described in Winkler's book [2005]. As a result, the memory requirement can be reduced by a factor of nearly 7, and the overall computation can be accelerated by efficiently computing convolution operations in graphics hardware. The downside is the metric's reduced prediction performance since second derivative Gaussian filters are not perceptually justified and our pilot implementation also indicated difficulties in calibration.

A limitation of our metric is the lack of a mechanism to model vi-

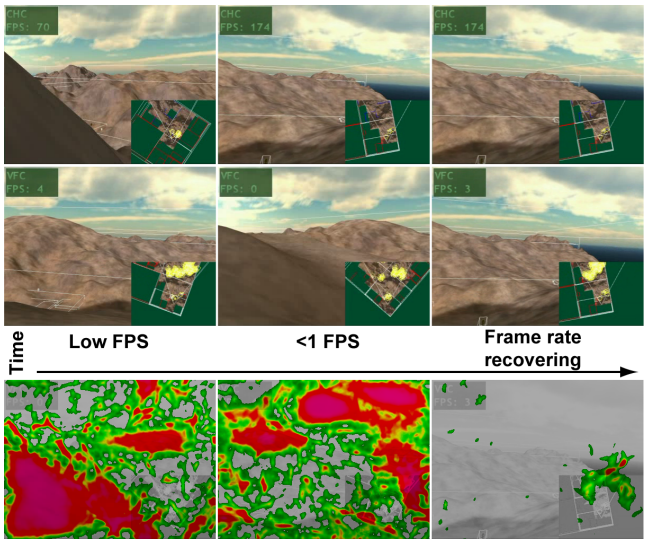

Figure 20: *The effect of dropped frames to perceived quality. One should note, however, that our method does not compensate for camera movements and assumes frames are perfectly aligned with each other.*

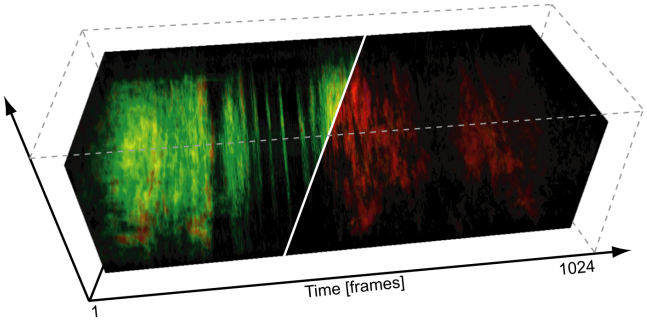

Figure 21: *An alternative 3D visualization. The left slice shows a volume rendering of the entire visible differences data. The right slice shows only the differences with detection probability above* 75% *where the locations of the missing frames along the time axis are better visible.*

sual attention. In the absence of either a computational model, or eye tracking data to predict the observer's gaze direction, our metric's predictions are conservative in the sense that the possibility of the observer focusing her attention to some other region than where the sought artifact appears is not considered. Another limitation of our metric is the requirement of a reference video for quality evaluation, which may not be available in some applications. No reference metrics, however, have limited utility since they are often geared toward detecting a single type of distortion, and are generally not as accurate as full reference metrics.

8 Conclusion

We presented a video quality metric specifically designed for Computer Graphics applications. Our method comprises an HVS model built with spatiotemporal components that are designed for HDR luminance levels. The capability of comparing video pairs with different dynamic ranges enables applications such as objective evaluation of HDR video compression and tone mapping, as well as comparison of different rendering methods and predicting the effect of dropped frames to perceived quality.

The validation of video quality metrics is often performed by comparing the metric responses to standard image quality databases. In the absence of such a collection of video pairs and corresponding spatial distortion maps comprising stimuli with different dynamic ranges and multitude of artifact types relevant to Computer Graphics, we created a modest data set for validation purposes. A future direction is to extend our initial effort to a standardized data set. Another possible extention to our work is the inclusion of color channels utilizing a color appearance model designed for HDR luminance levels. Temporal inverse tone mapping evaluation is a natural application area of our metric, but it was not included in this work since from the metric's point of view, the difference between forward and inverse tone mapping is merely swapping reference (HDR) and test (LDR) videos. Nevertheless, the metric's detection performance of application specific banding artifacts deserves further investigation.

Acknowledgements

Thanks to Robert Herzog for generating the rendered sequences, to Oliver Mattausch for view frustum culling sequences, and to Rafal Mantiuk for providing us with his HDR compression codes and helping us running it. Thanks to Jens Kerber for his help with volumetric visualizations, and to Makoto Okabe for editing and Glenn Lawyer for dubbing the supplemental video. Thanks to all the stuff members and students at MPI Informatik who participated in our experiments. Pisa HDR image and RNL HDR video courtesy of Paul Debevec.

References

AYDIN, T. O., MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2008. Dynamic range independent image quality assessment. In *Proc. of ACM SIGGRAPH*, vol. 27(3). Article 69.

BAVOIL, L., SAINZ, M., AND DIMITROV, R. 2008. Image-space horizon-based ambient occlusion. In *SIGGRAPH '08: ACM SIGGRAPH 2008 talks*, ACM, New York, NY, USA, 1–1.

BITTNER, J., WIMMER, M., PIRINGER, H., AND PURGATHOFER, W. 2004. Coherent hierarchical culling: Hardware occlusion queries made useful. *Computer Graphics Forum 23*, 3 (Sept.), 615–624. Proceedings EUROGRAPHICS 2004.

BOLIN, M., AND MEYER, G. 1998. A perceptually based adaptive sampling algorithm. In *Proc. of Siggraph'98*, 299–310.

DACHSBACHER, C., AND STAMMINGER, M. 2005. Reflective shadow maps. In *I3D '05: Proceedings of the 2005 symposium on Interactive 3D graphics and games*, ACM, New York, NY, USA, 203–231.

DALY, S. 1993. The Visible Differences Predictor: An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, MIT Press, A. B. Watson, Ed., 179–206.

DALY, S. J. 1998. Engineering observations from spatiovelocity and spatiotemporal visual models. SPIE, B. E. Rogowitz and T. N. Pappas, Eds., vol. 3299, 180–191.

DRAGO, F., MYSZKOWSKI, K., ANNEN, T., AND N.CHIBA. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum 22*, 3.

FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. In *SIGGRAPH*

'02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, ACM Press, 249–256.

FERWERDA, J., AND PELLACINI, F. 2003. Functional difference predictors (fdps): measuring meaningful image differences. In *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, 1388 – 1392 Vol.2.

FREDERICKSEN, R. E., H. R. F. 1998. Estimating multiple temporal mechanisms in human vision. In *Vision Research*, vol. 38, 1023–1040.

FREEMAN, W. T., AND ADELSON, E. H. 1991. The design and use of steerable filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 13*, 9, 891–906.

HERZOG, R., EISEMANN, E., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2010. Spatio-temporal upsampling on the GPU. In *I3D '10: Proceedings of the 2010 symposium on Interactive 3D graphics and games*, ACM, New York, NY, USA, 91–98.

ITU-T. 1999. Subjective video quality assessment methods for multimedia applications.

LINDH, P., AND VAN DEN BRANDEN LAMBRECHT, C. 1996. Efficient spatio-temporal decomposition for perceptual processing of video sequences. In *Proceedings of International Conference on Image Processing ICIP'96*, IEEE, vol. 3 of *Proc. of IEEE*, 331–334.

LUBIN, J. 1995. *Vision Models for Target Detection and Recognition*. World Scientific, ch. A Visual Discrimination Model for Imaging System Design and Evaluation, 245–283.

LUKIN, A. 2009. Improved visible differences predictor using a complex cortex transform. *GraphiCon*, 145–150.

MANTIUK, R., KRAWCZYK, G., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2004. Perception-motivated high dynamic range video encoding. *ACM Trans. Graph. 23*, 3, 733–741.

MANTIUK, R., DALY, S., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2005. Predicting visible differences in high dynamic range images - model and its calibration. In *Human Vision and Electronic Imaging X*, vol. 5666 of *SPIE Proceedings Series*, 204–214.

MASRY, M. A., AND HEMAMI, S. S. 2004. A metric for continuous quality evaluation of compressed video with severe distortions. *Signal Processing: Image Communication 19*, 2, 133 – 146.

MYSZKOWSKI, K., ROKITA, P., AND TAWARA, T. 2000. Perception-based fast rendering and antialiasing of walkthrough sequences. *IEEE Transactions on Visualization and Computer Graphics 6*, 4, 360–379.

MYSZKOWSKI, K., TAWARA, T., AKAMINE, H., AND SEIDEL, H.-P. 2001. Perception-guided global illumination solution for animation rendering. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA, 221–230.

PATTANAIK, S. N., TUMBLIN, J. E., YEE, H., AND GREENBERG, D. P. 2000. Time-dependent visual adaptation for fast realistic image display. In *Proc. of ACM SIGGRAPH 2000*, 47–54.

REEVES, W. T., SALESIN, D. H., AND COOK, R. L. 1987. Rendering antialiased shadows with depth maps. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA, 283–291.

RITSCHEL, T., GROSCH, T., AND SEIDEL, H.-P. 2009. Approximating dynamic global illumination in image space. In *I3D '09: Proceedings of the 2009 symposium on Interactive 3D graphics and games*, ACM, New York, NY, USA, 75–82.

RUSHMEIER, H., WARD, G., PIATKO, C., SANDERS, P., AND RUST, B. 1995. Comparing real and synthetic images: some ideas about metrics. In *Rendering Techniques '95*, Springer, P. Hanrahan and W. Purgathofer, Eds., 82–91.

SAMPAT, M. P., WANG, Z., GUPTA, S., BOVIK, A. C., AND MARKEY, M. K. 2009. Complex wavelet structural similarity: A new image similarity index. *Image Processing, IEEE Transactions on 18*, 11 (Nov.), 2385–2401.

SCHWARZ, M., AND STAMMINGER, M. 2009. On predicting visual popping in dynamic scenes. In *APGV '09: Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, ACM, New York, NY, USA, 93–100.

SESHADRINATHAN, K., AND BOVIK, A. 2007. A structural similarity metric for video based on motion models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, I–869–I–872.

SESHADRINATHAN, K., AND BOVIK, A. C. 2010. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on 19*, 2 (Feb.), 335– 350.

VAN DEN BRANDEN LAMBRECHT, C., AND VERSCHEURE, O. 1996. Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System. In *IS&T/SPIE*.

VAN DEN BRANDEN LAMBRECHT, C., COSTANTINI, D., SICURANZA, G., AND KUNT, M. 1999. Quality assessment of motion rendition in video coding. *Circuits and Systems for Video Technology, IEEE Transactions on 9*, 5 (Aug), 766–782.

WANDELL, B. A. 1995. *Foundations of Vision*. Sinauer Associates, Inc.

WANG, Z., AND SIMONCELLI, E. 2005. Translation insensitive image similarity in complex wavelet domain. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 2, 573–576.

WATSON, A. B., AND MALO, J. 2002. Video quality measures based on the standard spatial observer. In *ICIP (3)*, 41–44.

WATSON, A. B., HU, J., AND III, J. F. M. 2001. DVQ: A digital video quality metric based on human vision. *Journal of Electronic Imaging 10*, 20–29.

WATSON, A. B. 1986. Temporal sensitivity. In *Handbook of Perception and Human Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds. John Wiley and Sons, New York, 6–1–6–43.

WATSON, A. 1987. The Cortex transform: rapid computation of simulated neural images. *Comp. Vision Graphics and Image Processing 39*, 311–327.

WINKLER, S. 1999. A perceptual distortion metric for digital color video. In *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging*, IEEE, vol. 3644 of *Controlling Chaos and Bifurcations in Engineering Systems*, 175–184.

WINKLER, S. 2005. *Digital Video Quality: Vision Models and Metrics*. Wiley.

YEE, H., PATTANAIK, S., AND GREENBERG, D. P. 2001. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Trans. Graph. 20*, 1, 39–65.