HDR Video Metrics

Martin Čadík, Tunç Ozan Aydın

CPhoto@FIT, http://cphoto.fit.vutbr.cz,
Faculty of Information Technology,
Brno University of Technology,
Brno, Czech Republic

Abstract

In this chapter we present an introduction to HDR image and video quality assessment fields. We discuss full-, no-, and reduced-reference metrics, including perceptually motivated methods. We describe two existing full-reference HDR video quality metrics in detail. Furthermore, we introduce the emerging field of data-driven metrics. Finally, we conclude with the outlook of future development and research.

Keywords: Video Quality Assessment, Image Quality Assessment, Image Quality Metrics, Video Quality Metrics, Objective Quality Assessment, HDR, Tone Mapping

Email address: cadikm@centrum.cz, tuncozanaydin@gmail.com (Martin Čadík, Tunç Ozan Aydın)

High Dynamic Range Video, Elsevier, 2016, ISBN 9780128094778.

1. Introduction

In this chapter we present an introduction to HDR image and video quality assessment fields. We discuss full-, no-, and reduced-reference metrics, including perceptually motivated methods. We describe two existing full-reference HDR video quality metrics in detail. Furthermore, we introduce the emerging field of data-driven metrics. Finally, we conclude with the outlook of future development and research.

2. Image and Video Quality Assessment

The goal of image and video quality assessment (IQA, VQA) is to computationally predict human perception of image and video quality. Practical evidence shows [1, 2] that numerical distortion metrics, like root mean squared error (RMSE), are often not adequate for the comparison of images, because they poorly predict the differences between the images as perceived by a human observer. To solve this problem properly, various *image and video quality metrics* (IQM, VQM) have been proposed [2]. Image quality metrics traditionally comprise a computational human visual system (HVS) model to correctly predict image difference as a human would perceive it, be it a bottom-up [3], or a top-down approach [4]. Please refer to vision science textbooks [5] for an in-depth treatment of human visual perception, and on HVS measurements related to masking, adaptation, contrast sensitivity, etc.

Image and video quality assessment is practical in various applications. The main applications of IQA lie in the areas of image quality monitoring (e.g. in lossy image compression), benchmarking of imaging applications, and optimizing algorithms by tuning their parameter settings. Furthermore, image quality metrics have also been successfully applied to image database retrievals, or evaluation of the perceptual impact of different computer graphics and vision algorithms.

In the following text we will survey existing standard dynamic range (SDR) quality assessment approaches, while the only two existing metrics designed specifically for HDR video processing will be described in sections 3 and 4 in detail.

2.1. Full-reference Metrics

Full-reference image and video quality metrics are based on measuring the errors (signal differences) between a distorted image and the reference image. The aim is to quantify the errors in a way that simulates human visual error sensitivity. Video quality assessment is often inspired by the ideas from the more developed image quality assessment field. A great variety of SDR image quality metrics have been proposed in the literature [1, 2]. Traditionally, image quality metrics focus on near-threshold detection [6], supra-threshold discrimination [7], or functional differences [8].

Video metrics usually extend image quality 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) [9] 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 [10]. In later work targeted specifically to assess the quality of MPEG-2 compressed videos [11], 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 [12], where they use finite impulse response filters for temporal decomposition.

Watson et al. [13] proposed 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 simple Standard Spatial Observer (SSO) based method [14], which, on the Video Quality Experts Group data set, is shown to make as accurate predictions as more complex metrics. Winkler [15, 16] 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.

A video metric V-SSIM proposed by Seshadrinathan and Bovik [17] is an extension to the image quality metric called Complex Wavelet Structural Similarity Index (CW-SSIM [18, 19]) to account for motion in video sequences. The technique 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. The same authors published the MOVIE index in a follow-up work [20], 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).

2.2. No-reference and Reduced-reference Metrics

The main issue in developing *no-reference (NR) metrics* is the absence of a non-distorted reference image or some features representing it. Common approaches to compensate for this are (1) modeling distortion-specific characteristics, (2) using natural scene statistics, and (3) employing learning based classification methods.

Distortion-specific NR methods capitalize on the knowledge of artifact type and its unique characteristics [21, Ch. 3]. Examples include metrics for detecting blockiness due to lossy JPEG and MPEG compression and ringing at strong contrast edges [22], blurriness due to high frequency coefficients suppression

[23, 24], banding (false contouring) at low gradient regions due to the excessive quantization [25]. There are some attempts of building more general NR quality metrics, which evaluate a combined contribution of individually estimated image features such as sharpness, contrast, noise, clipping, ringing, and blocking artifacts [21, Ch. 10]. The contribution of all features including their simple interactions is summed up with weights derived through fitting to subjective data.

Natural scene statistics [26] derived from artifact-free images can be helpful in detecting artifacts. Sheikh et al. show that noise, blurriness, and quantization can be identified as deviations from these statistics [27].

Image features extracted from distorted and non-distorted images are used for training machine learning techniques such as support vector machines (SVM) or neural networks. Moorthy and Bovik [28] use generalized Gaussian distribution (GGD) to parametrize wavelet subband coefficients and create 18-D feature vector (3 scales $\times$ 3 orientations $\times$ 2 GGD parameters), which is used to train an SVM classifier based on perceptually calibrated distortion examples from the LIVE IQA database. The classifier discriminates between five types of mostly compression-related distortions and estimates their magnitude. Saad et al. [29] train a statistical model to detect distortions in DCT-based contrast and structure features.

Reduced-reference metrics may be beneficial in video compression or transmission monitoring [30, 31, 32, 33], where the bandwidth is limited. The challenge is to select a representative set of features, which are extracted from an undistorted signal and transmitted along with the possibly distorted image or video. In their pioneering work, Webster et al. [34] used localized spatial and temporal activity channels for this purpose. Later on, Redi et al. [35] identified the *color correlograms* as suitable feature descriptors for analysis of alterations in the color distribution as a result of distortions.

3. DRI-VQM

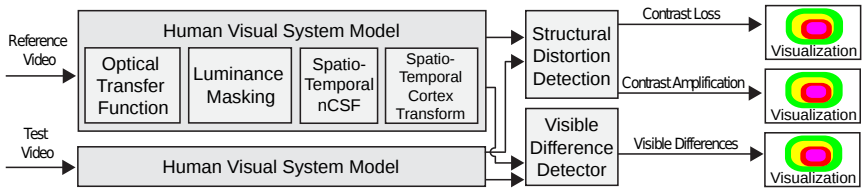

Figure 1: Data flow diagram of DRI-VQM. See text for details.

The recent proliferation of High Dynamic Range (HDR) Imaging forces video quality assessment metrics to be accurate in extended luminance ranges. This requirement limits the use of legacy video quality metrics designed for detecting compression artifacts in standard dynamic range (SDR) 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 SDR. DRI-VQM is a video quality metric that is designed specifically for these recently emerged practical problems.

DRI-VQM utilizes an HDR capable human visual system (HVS) model that accounts for both major spatial and temporal aspects of the visual system, and employs a pair of dynamic range independent distortion measures *contrast loss* and *amplification* introduced in its counterpart for images, DRI-IQM [36]. DRI-VQM also computes the *visible differences* between reference and test videos similar to conventional video quality metrics. In most visual computing and video quality assessment applications the main concern is often the existence of visible artifacts rather than the magnitude of visibility. Methods that produce clearly visible artifacts are often not useful in practice. Consequently DRI-VQM's HVS model trades off supra-threshold precision for accuracy near the detection threshold.

The computational steps of DRI-VQM are summarized in Figure 1. 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 SDR videos should be inverse gamma corrected and converted to display luminance (In all examples we assumed a hypothetical display 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 multi-channel local contrast at each visual channel, where the first step is to model the nonlinear response of the photoreceptors to luminance, namely **Luminance adaptation**. In DRI-VQM we apply the non-linearity 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.

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 2c 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 [6] (Figure 2a) 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 [37]. In the absence of eye tracking data DRI-VQM assumes that the observer's gaze is fixed, but also the drift movement is present. Accordingly, a minimum retinal velocity is set as follows:

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

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 2d) by first computing a *Luminance Modulation Factor* (Figure 2b) 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:

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

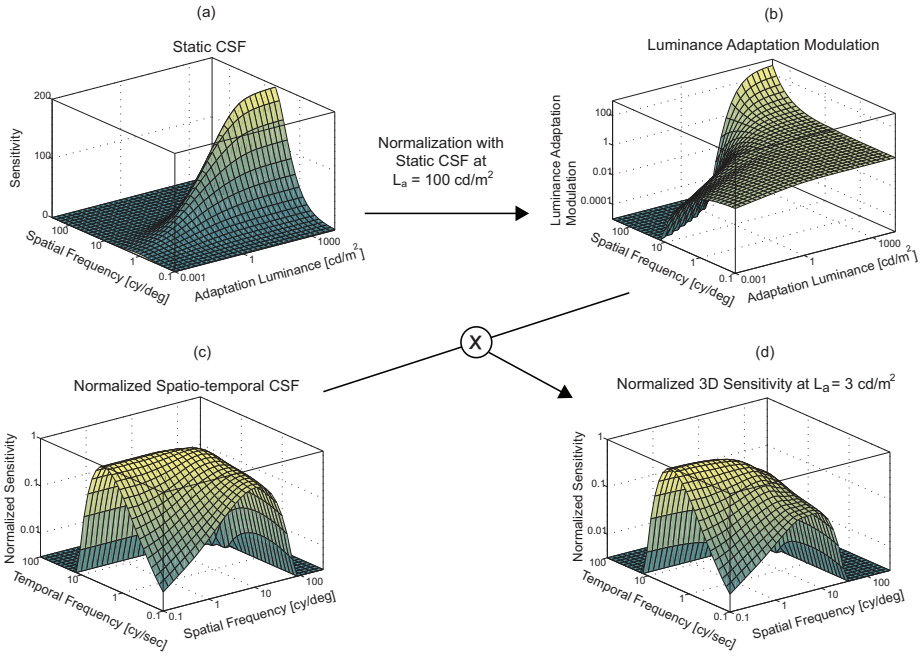

Figure 2: Computation of the CSF^{3D}. The static $CSF^S(\rho, L_a)$ (a) is divided to $CSF^S(\rho, L_a = 100cd/m^2)$ to obtain scaling coefficients (b) that account for luminance adaptation in CSF^{3D}. The specific adaptation level is chosen to reflect the conditions where the spatiotemporal CSF^T 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).

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, estimating luminance adaptation using a scaling factor is better than the alternatives that involve an approximation by linear separation of spatial and temporal frequencies. The effect of luminance adaptation to spatiotemporal contrast sensitivity is approximately linear except for very low temporal frequencies [38, 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 DRI-VQM extends the **Cortex Transform** [39] 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 3-a, are given as Equation 3 and its second derivative, respectively [16]:

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

The corresponding frequency domain filters are computed by applying the Fourier transform to both impulse responses and are shown in Figure 3-b.

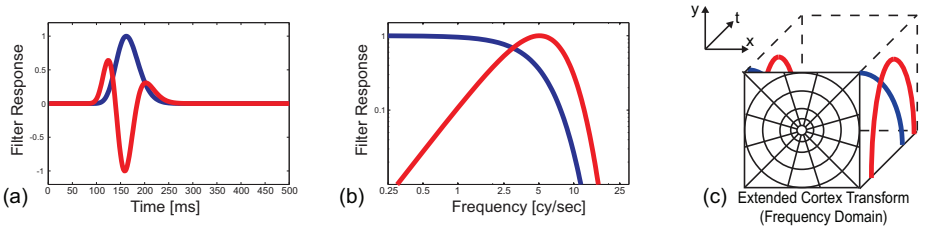

Figure 3: Impulse (a) and frequency (b) responses of the transient (red) and sustained (blue) temporal channels. The frequency responses comprise the extended 3D Cortex Transform's channels in temporal dimension (c).

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

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

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 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:

$$P(C) = 1 - \exp(-|C|^3). \tag{4}$$

We compute the probability of detecting a **visible difference** between videos ($P(C_{tst} - C_{ref})$), as well as two dynamic range independent distortion measures from individual detection probabilities of the contrast in visual channels [36]. The per-channel dynamic range independent distortion measures are defined as follows:

- **Contrast Loss** $\left(P^{k,l,m}_{\searrow} = P(C^{k,l,m}_{ref})(1 - P(C^{k,l,m}_{tst})\right)$
- **Contrast Amplification** $\left(P^{k,l,m}_{\nearrow} = P(C^{k,l,m}_{tst})(1 - P(C^{k,l,m}_{ref})\right)$.

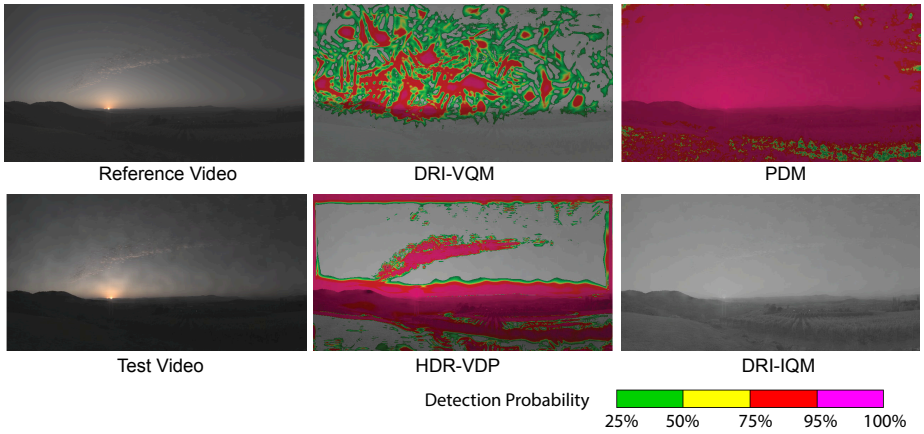

Figure 4: Comparison of DRI-VQM with other video and image quality metrics, which fail to predict the visibility of the noise pattern present in the test video.

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. Detection probabilities of each type of distortions are then combined using a standard probability summation function:

$$\hat{P}_{\Delta|\searrow|\nearrow} = 1 - \prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{m=1}^{M}\left(1 - P^{k,l,m}_{\Delta|\searrow|\nearrow}\right). \tag{5}$$

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.

The implementation of DRI-VQM video metric is publicly available online (`http://metrics.mpi-inf.mpg.de/`) along with other metrics.

4. HDR-VQM

HDR-VQM, an alternative to DRI-VQM described in Section 3, has been proposed recently [40]. An overview of the HDR-VQM metric is shown in Figure 6. Similarly to DRI-VQM, the HDR-VQM is a full-reference HDR video

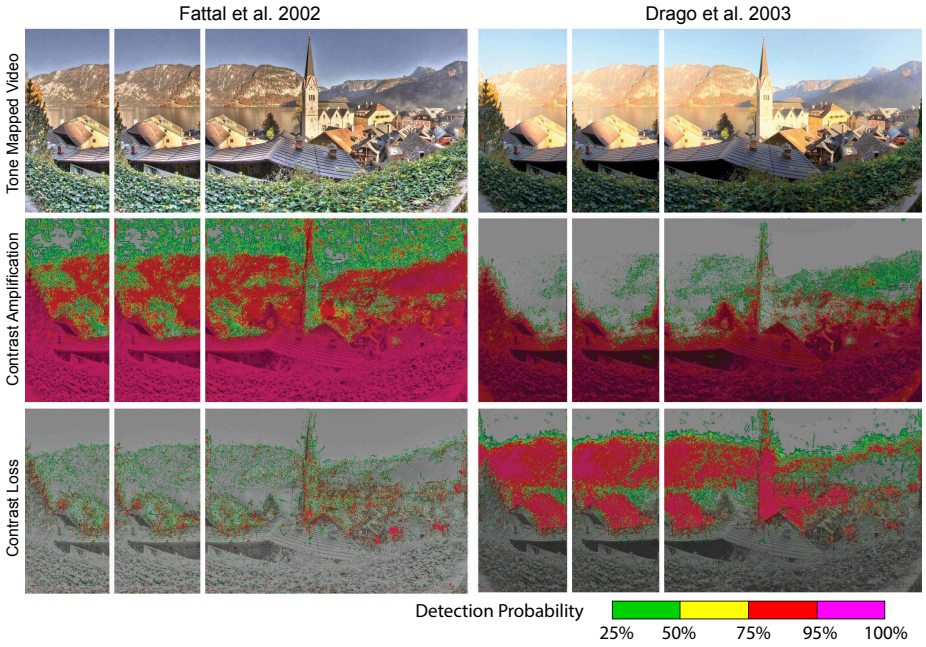

Figure 5: In-context visualization of the contrast loss and amplification of SDR videos obtained by Fattal and Drago tone mapping operators with respect to the reference HDR video.

quality metric hence their building blocks are similar as well. The method is based on signal pre-processing, transformation, frequency based decomposition and subsequent spatio-temporal pooling, as described in more detail below. However, the main difference resides in the application area. HDR-VQM targets the signal processing, video transmission and related fields, where the distortion of the signal is often considerable and the information about the *overall video* quality is thus an expected and sufficient measure. Accordingly, HDR-VQM aims to predict human perception of the *supra-threshold* video distortions, which are then pooled to a single number, a measure of an overall video quality.

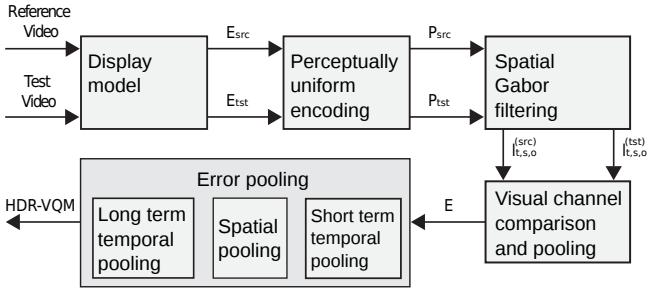

Figure 6: Data flow diagram of HDR-VQM. See text for details.

4.1. Transformation into emitted luminance

First, the input videos are transformed into the luminance values emitted by the display device. This is a difficult problem, because the HDR values encoded in the HDR video are often not calibrated (i.e. relative), and thus they are merely proportional to the input luminance. Moreover, the accurate display processing model is usually unknown. Instead, the authors of HDR-VQM adopt a simple approximation using a scaling factor as follows. The input HDR videos are normalized by the maximum of the mean of top 5% HDR values of all the frames in the video sequence. A clipping function is finally applied to mimic the physical limitations of the display. This way, the values of the emitted luminance E fit in the range given by the black point of the display and the highest displayable luminance. The values outside this range are saturated, representing the information loss due to the display device.

4.2. From emitted to perceived luminance

The second step approximates the human perception P of the emitted luminance E, which is known to be nonlinear [5], approximately logarithmic. To model this behavior in HDR-VQM, the perceptually uniform (PU) encoding proposed by Aydın et al. [41] was adopted, see Figure 7. The central idea of the PU encoding is to make differentials of the curve proportional to the luminance detection thresholds. The PU encoding is expected to model the HVS in a better way than a simple logarithmic function, yet still it is only a crude approximation of the HVS luminance response. However, the PU encoding may be implemented efficiently as a look-up table operation (available from `http://resources.mpi-inf.mpg.de/hdr/fulldr_extension/`).

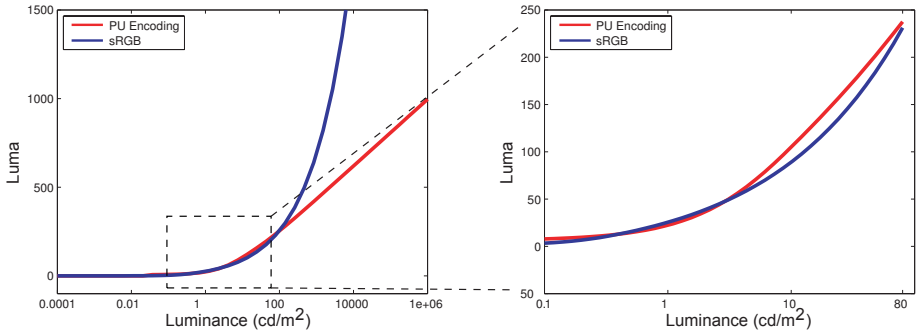

Figure 7: Perceptually uniform (PU) encoding is backward-compatible with the sRGB nonlinearity. The curve is shown along the entire dynamic range (left), and only within the operating range of sRGB (right).

4.3. Decomposition into visual channels

Similarly to DRI-VQM, the perceived luminance P is subsequently decomposed into visual channels. However, the implemented decomposition distinguishes only spatial frequencies and orientations, leaving temporal processing to

the later pooling stage for efficiency. Consequently, the spatio-temporal contrast sensitivity (CSF) of human visual system can not be modeled. More specifically, the employed decomposition is based on log-Gabor filters [42] implemented in the frequency domain. This way, the reference and distorted videos are decomposed into visual channels (subbands) $\{l_{t,s,o}\}$, where $s = 1, 2, ..., N_{scale}$ is the total number of scales, $o = 1, 2, ..., N_{orient}$ is the total number of orientations, and $t = 1, 2, ..., F$ is the number of frames in the sequence.

The error in each channel is then computed per frame using a simple bounded measure as follows:

$$E_{t,s,o} = \frac{2l_{t,s,o}^{(src)} l_{t,s,o}^{(dst)} + k}{(l_{t,s,o}^{(src)})^2 + (l_{t,s,o}^{(dst)})^2 + k},$$

where k is a small constant to avoid division by zero.

4.4. Pooling

First, a simple error pooling across scales and orientations is performed. This neglects contrast sensitivity of the human visual system (CSF), which is essential to model near-threshold sensitivity. Assuming supra-threshold distortions, the pooling boils down to a simple equal weighting as follows: $E_t = \frac{1}{N_{scale} N_{orient}} \sum_{s=1}^{N_{scale}} \sum_{o=1}^{N_{orient}} E_{t,s,o}$, where E_t is a per-frame distortion map. Please notice that no temporal processing has been involved so far, therefore the resulting distortion video $E = \{E_t\}_{t=1}^{F}$ is equivalent to computing an *image* quality metric separately for each video frame t. On the other hand, DRI-VQM described above involves spatio-*temporal* decomposition followed by the spatio-*temporal* CSF filtering. The distortion video produced by DRI-VQM therefore accounts for temporal behavior of the human visual system.

The subsequent spatio-temporal HDR-VQM pooling step is an interesting way of modeling temporal perception, which has not been considered in the previous steps of the algorithm. Motivated by the alternations in visual fixations, the distortion video E is first divided into non-overlapping short-term tubes (channels) ST defined by a 3D region $x \times y \times z$. The spatial extent of ST regions ($x \times y$) is given by the viewing distance, the central angle of the visual field in the fovea and the display resolution. The temporal dimension z is defined by average eye fixation time, and it is set to 300-500ms in HDR-VQM. Consequently, the *short term temporal pooling* is performed by computing the standard deviation of each ST tube. This results in spatio-temporal subband error frames $\{ST_{v,t_s}\}_{t_s=1}^{F/z}$, where v represents the spatial coordinates. Finally, the *spatial and long term temporal pooling* is performed to yield the global video quality score in a simple way as follows. First, the subband error frames ST_{v,t_s} are pooled spatially resulting in a time series of short term quality scores. Finally, in a long term temporal pooling, the time series are fused to a single number denoting *overall video quality*:

$$\text{HDR-VQM} = \frac{1}{|t_s \in L_p||v \in L_p|} \sum_{t_s \in L_p} \sum_{v \in L_p} ST_{v,t_s},$$

where L_p denotes the set with lowest $p\%$ values[1], and $|.|$ is a cardinality of the set. The pooling factor p is set to 5%, however according to the authors, varying p between 5% to 50% does not significantly change the prediction accuracy. It should be noted that contrary to DRI-VQM, HDR-VQM does not explicitly model spatio-temporal *masking* effects.

5. Data-driven Metrics

Even though knowledge about the human visual system (HVS) is continuously expanded, many unanswered questions and unverified hypotheses still remain. On that account, we are quite far from having an accurate bottom-up model of the HVS. Therefore, additionally to the bottom-up approaches shown above, top-down *data-driven approaches* based on machine learning are starting to emerge. Machine learning techniques have recently gained a lot of popularity and attention in many research areas. For such methods, it is of crucial importance to provide a sufficient amount of training data. Unfortunately, not many usable datasets exhibiting localized distortion maps measured on human subjects are available. Accordingly, the possibilities of data-driven approaches are currently being explored on simpler, *image* quality assessment task.

More specifically, two experiments [43] were performed where observers used a brush-painting interface to directly mark distorted image regions in the presence and absence of a high-quality reference image. The resulting *per-pixel image-quality datasets* enabled a thorough evaluation of existing full-reference metrics and the development of new machine learning-based metrics.

Specifically, the datasets were utilized to develop a *Learning-based Predictor of Localized Distortions (LPLD)* [44]. LPLD is a *full-reference metric* for synthetic images. The key element of the metric is a carefully designed set of features, which generalize over distortion types, image content, and superposition of multiple distortions in a single image. Additionally, two new datasets to validate this metric were created and made publicly available (`http://resources.mpi-inf.mpg.de/hdr/metric/`): a continuous range of basic distortions encapsulated in a few images, and the distortion saliency maps captured in the eye tracking experiment. The distortion maps are useful to benchmark existing and future metrics and associated saliency maps could be used, for instance, in perceptual studies of human visual attention.

Finally, a data-driven *no-reference image quality metric* for synthetic images called *NoRM* [45] was proposed. NoRM uses a supervised learning algorithm to predict a perceptual distortion map, which measures the probability of noticing the local distortions on the pixel-level. The proposed metric achieves prediction performance comparable to full-reference metrics. Besides the machine learning machinery, the quality of the results of NoRM is owed to rendering-specific features extracted from the depth map and the surface-material information.

[1] Both short term spatial and long term temporal pooling is performed only over the lowest $p\%$ values.

6. Outlook and Future Work

Despite many years of active research on image and video quality assessment, the developed metrics are often still far from being comparable to human observers. Existing universal metrics are not mature and robust enough to stand in all scenarios. However, to overcome this issue, one may develop specialized metrics tailored specifically to the particular problem. Recent examples of such metrics include the quality predictor for image completion [46], or similarity measure for illustration style [47]. Furthermore, measuring vaguely defined quantities like interestingness of images [48] or aesthetic and beauty [49, 50] may be also feasible, perhaps thanks to the machine learning algorithms. Finally, the emerging area of multispectral image and video comparison [51] remains currently almost unexplored.

Acknowledgements

This work was supported by SoMoPro II grant (financial contribution from the EU 7 FP People Programme Marie Curie Actions, REA 291782, and from the South Moravian Region). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

References

[1] Z. Wang, A. C. Bovik, Modern image quality assessment, Synthesis Lectures on Image, Video, and Multimedia Processing 2 (1) (2006) 1–156.

[2] H. R. Wu, K. R. Rao, Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications), CRC Press, Inc., Boca Raton, FL, USA, 2005.

[3] R. Mantiuk, K. J. Kim, A. G. Rempel, W. Heidrich, HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions, ACM Transactions on Graphics (Proc. of SIGGRAPH) (2011) 40:1–40:14.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, S. Member, E. P. Simoncelli, S. Member, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.

[5] S. E. Palmer, Vision science – photons to phenomenology, 3rd Edition, The MIT Press, Cambridge, 2002.

[6] S. Daly, The Visible Differences Predictor: An algorithm for the assessment of image fidelity, in: Digital Images and Human Vision, MIT Press, 1993, pp. 179–206.

[7] J. Lubin, Vision Models for Target Detection and Recognition, World Scientific, 1995, Ch. A Visual Discrimination Model for Imaging System Design and Evaluation, pp. 245–283.

[8] J. Ferwerda, F. Pellacini, 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, 2003, pp. 1388 – 1392 Vol.2. doi:10.1109/ACSSC.2003.1292214.

[9] C. van den Branden Lambrecht, O. Verscheure, Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System, in: IS&T/SPIE, 1996.

[10] P. Lindh, C. van den Branden Lambrecht, Efficient spatio-temporal decomposition for perceptual processing of video sequences, in: Proceedings of International Conference on Image Processing ICIP'96, Vol. 3 of Proc. of IEEE, IEEE, 1996, pp. 331–334.

[11] C. van den Branden Lambrecht, D. Costantini, G. Sicuranza, M. Kunt, Quality assessment of motion rendition in video coding, Circuits and Systems for Video Technology, IEEE Transactions on 9 (5) (1999) 766–782. doi:10.1109/76.780365.

[12] M. A. Masry, S. S. Hemami, A metric for continuous quality evaluation of compressed video with severe distortions, Signal Processing: Image Communication 19 (2) (2004) 133 – 146. doi:DOI:10.1016/j.image.2003.08.001.
URL http://www.sciencedirect.com/science/article/B6V08-49FPGN9-2/2/87da5642b73b1d703797de2099845d6b

[13] A. B. Watson, J. Hu, J. F. M. Iii, DVQ: A digital video quality metric based on human vision, Journal of Electronic Imaging 10 (2001) 20–29.

[14] A. B. Watson, J. Malo, Video quality measures based on the standard spatial observer, in: ICIP (3), 2002, pp. 41–44.

[15] S. Winkler, A perceptual distortion metric for digital color video, in: Proceedings of the SPIE Conference on Human Vision and Electronic Imaging, Vol. 3644 of Controlling Chaos and Bifurcations in Engineering Systems, IEEE, 1999, pp. 175–184.

[16] S. Winkler, Digital Video Quality: Vision Models and Metrics, Wiley, 2005.

[17] K. Seshadrinathan, A. Bovik, 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, 2007, pp. I–869–I–872. doi:10.1109/ICASSP.2007.366046.

[18] Z. Wang, E. Simoncelli, Translation insensitive image similarity in complex wavelet domain, in: Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on, Vol. 2, 2005, pp. 573–576.

[19] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, M. K. Markey, Complex wavelet structural similarity: A new image similarity index, Image Processing, IEEE Transactions on 18 (11) (2009) 2385–2401. doi:10.1109/TIP.2009.2025923.

[20] K. Seshadrinathan, A. C. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, Image Processing, IEEE Transactions on 19 (2) (2010) 335– 350.

[21] H. Wu, K. Rao, Digital Video Image Quality and Perceptual Coding, CRC Press, 2005.

[22] Z. Wang, A. C. Bovik, Modern Image Quality Assessment, Morgan & Claypool Publishers, 2006.

[23] C. Chen, W. Chen, J. A. Bloom, A universal reference-free blurriness measure, in: SPIE vol. 7867, 2011. doi:10.1117/12.872477.

[24] H. Liu, I. Heynderickx, Issues in the design of a no-reference metric for perceived blur, in: SPIE vol. 7867, 2011. doi:10.1117/12.873277.

[25] S. Daly, X. Feng, Decontouring: Prevention and removal of false contour artifacts, in: Proc. of Human Vision and Electronic Imaging IX, SPIE, vol. 5292, 2004, pp. 130–149.

[26] E. P. Simoncelli, Statistical modeling of photographic images, in: A. C. Bovik (Ed.), Handbook of Image and Video Processing, Academic Press, Inc., 2005, pp. 431–441.

[27] H. Sheikh, A. Bovik, L. Cormack, No-reference quality assessment using natural scene statistics: JPEG2000, IEEE Trans. on Image Processing 14 (11) (2005) 1918 –1927. doi:10.1109/TIP.2005.854492.

[28] A. Moorthy, A. Bovik, A two-step framework for constructing blind image quality indices, IEEE Signal Processing Letters 17 (5) (2010) 513 –516. doi:10.1109/LSP.2010.2043888.

[29] M. Saad, A. Bovik, C. Charrier, A DCT statistics-based blind image quality index, IEEE Signal Processing Letters 17 (6) (2010) 583 –586. doi:10.1109/LSP.2010.2045550.

[30] T. Oelbaum, K. Diepold, A reduced reference video quality metric for avc/h.264, in: Signal Processing Conference, 2007 15th European, 2007, pp. 1265–1269.

[31] L. Ma, S. Li, K. N. Ngan, Reduced-reference video quality assessment of compressed video sequences, Circuits and Systems for Video Technology, IEEE Transactions on 22 (10) (2012) 1441–1456. doi:10.1109/TCSVT.2012.2202049.

[32] M. Martini, B. Villarini, F. Fiorucci, A reduced-reference perceptual image and video quality metric based on edge preservation, EURASIP Journal on Advances in Signal Processing 2012 (1). doi:10.1186/1687-6180-2012-66.
URL http://dx.doi.org/10.1186/1687-6180-2012-66

[33] R. Soundararajan, A. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, Circuits and Systems for Video Technology, IEEE Transactions on 23 (4) (2013) 684–694. doi:10.1109/TCSVT.2012.2214933.

[34] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, S. Wolf, An objective video quality assessment system based on human perception, in: in SPIE Human Vision, Visual Processing, and Digital Display IV, 1993, pp. 15–26.

[35] J. a. Redi, P. Gastaldo, I. Heynderickx, R. Zunino, Color Distribution Information for the Reduced-Reference Assessment of Perceived Image Quality, IEEE Transactions on Circuits and Systems for Video Technology 20 (12) (2010) 1757–1769. doi:10.1109/TCSVT.2010.2087456.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5604294

[36] T. O. Aydın, R. Mantiuk, K. Myszkowski, H.-P. Seidel, Dynamic range independent image quality assessment, in: Proc. of ACM SIGGRAPH, Vol. 27(3), 2008, article 69.

[37] S. Daly, Engineering observations from spatiovelocity and spatiotemporal visual models, in: Proc. of SPIE: Human Vision and Electronic Imaging III, Vol. 3299, 1998.

[38] B. A. Wandell, Foundations of Vision, Sinauer Associates, Inc., 1995.

[39] A. B. Watson, The Cortex transform: rapid computation of simulated neural images, Comp. Vision Graphics and Image Processing 39 (1987) 311–327.

[40] M. Narwaria, M. P. D. Silva, P. L. Callet, HDR-VQM: An objective quality measure for high dynamic range video, Signal Processing: Image Communication 35 (2015) 46 – 60. doi:http://dx.doi.org/10.1016/j.image.2015.04.009.
URL http://www.sciencedirect.com/science/article/pii/S0923596515000703

[41] T. O. Aydın, R. Mantiuk, K. Myszkowski, H.-P. Seidel, Extending quality metrics to full luminance range images, in: Proc. of SPIE: Human Vision and Electronic Imaging XIII, Vol. 6806, 2008.

[42] D. J. Field, Relations between the statistics of natural images and the response properties of cortical cells, J. Opt. Soc. Am. A 4 (1987) 2379–2394.

[43] M. Čadík, R. Herzog, R. Mantiuk, K. Myszkowski, H.-P. Seidel, New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts, in: ACM Transactions on Graphics (Proc. of SIGGRAPH Asia), Vol. 31, ACM, 2012, pp. 1–10.

[44] M. Čadík, R. Herzog, R. Mantiuk, R. Mantiuk, K. Myszkowski, H. Seidel, Learning to predict localized distortions in rendered images, Computer Graphics Forum 32 (7) (2013) 401–410. doi:10.1111/cgf.12248.

[45] R. Herzog, M. Čadík, T. O. Aydın, K. I. Kim, K. Myszkowski, H.-P. Seidel, NoRM: no-reference image quality metric for realistic image synthesis, Computer Graphics Forum 31 (2) (2012) 545–554. doi:10.1111/j.1467-8659.2012.03055.x.

[46] J. Kopf, W. Kienzle, S. Drucker, S. B. Kang, Quality prediction for image completion, ACM Transactions on Graphics 31 (6) (2012) 131:1–131:8. doi:10.1145/2366145.2366150.
URL http://doi.acm.org/10.1145/2366145.2366150

[47] E. Garces, A. Agarwala, D. Gutierrez, A. Hertzmann, A similarity measure for illustration style, ACM Transactions on Graphics 33 (4) (2014) 93:1–93:9. doi:10.1145/2601097.2601131.
URL http://doi.acm.org/10.1145/2601097.2601131

[48] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, L. Van Gool, The interestingness of images, in: The IEEE International Conference on Computer Vision (ICCV), 2013.

[49] L. Marchesotti, N. Murray, F. Perronnin, Discovering beautiful attributes for aesthetic image analysis, Int. J. Comput. Vision 113 (3) (2015) 246–266. doi:10.1007/s11263-014-0789-2.
URL http://dx.doi.org/10.1007/s11263-014-0789-2

[50] T. O. Aydın, A. Smolic, M. Gross, Automated aesthetic analysis of photographic images, IEEE Transactions on Visualization and Computer Graphics 21 (1).

[51] S. Le Moan, P. Urban, Image-difference prediction: From color to spectral, Image Processing, IEEE Transactions on 23 (5) (2014) 2058–2068. doi:10.1109/TIP.2014.2311373.