SETTING VIDEO QUALITY & PERFORMANCE TARGETS FOR HDR AND WCG VIDEO SERVICES

SEAN T. MCCARTHY

TABLE OF CONTENTS

INTRODUCTION

High Dynamic Range (HDR) and Wide Color Gamut (WCG) can have a big positive impact on a viewer by creating a more convincing and compelling sense of light than has ever before been possible in television. A recent scientific study[1] with professional-quality Standard Dynamic Range (SDR) and HDR videos found that viewers prefer HDR over SDR by a large margin. Moreover, the study also showed that the margin of preference for HDR increased with increasing peak luminance.

What happens though to a viewer's quality of experience when pristine high quality HDR content is compressed for distribution? What happens when HDR WCG content is converted to SDR content to support legacy displays and consumer set-top boxes? Do distortions and compression artifacts become more noticeable in HDR? Does processed HDR lose some of its sparkle and become less discernible from ordinary SDR?

Video quality is easy to recognize by eye, but putting a number on video quality is often more problematic. For HDR & WCG the problem is even harder. HDR & WCG are so perceptually potent because even relatively infrequent features such as specular reflections and saturated colors can engage a viewer's attention fully. Yet, well-known video-quality scoring methods, such as peak signal-to-noise ratio (PSNR) and the Structural SIMilarity metric[2] (SSIM), could lead to wrong conclusions when applied to the perceptual outliers in HDR WCG video. Without good video-quality metrics, cable operators cannot make informed decisions when setting bitrate and video-quality performance targets, nor when choosing technology partners for HDR WCG services.

We need a way of quantifying distortions introduced during HDR WCG video processing that takes into account the wide luminance range of HDR video as well as the localized highlights, deep darks, and saturated colors that give HDR WCG its special appeal[3].

This paper introduces easy-to-calculate quantitative methods to provide cable operators with video-quality data that can be used to make operational, technological, and product decisions. Specifically, it presents methods to report the level of overall distortions in processed video as well as the specific distortions associated with perceptually important bright & dark HDR features and textures with respect to both luma and chroma components. The paper's objective is to show data and analysis that illustrates how quantifying HDR WCG video distortion can be made accurate, actionable, and practical, particularly when MSOs consider the various trade-offs between bandwidth, technology options, and the viewer's experience.

Quantifying HDR WCG Video Quality & Distortions

The best way to quantify video quality and viewer preference is to perform subjective testing using established techniques and existing international standards such as ITU-R

BT.500[4] and ITU-T P.910[5]; but subjective testing is too slow to be practical in most situations. Instead, a number of objective video quality assessment techniques and metrics have been developed over the decades[6]. Objective video quality assessment relies on computer algorithms that can be inserted into production and distribution workflows to provide actionable information. Some video quality algorithms, such as PSNR, are very simple, but do not correlate well with subjective scores[7,8]. Others are very sophisticated and include models of the human visual system. Such metrics do a better job of predicting subjective results, but can suffer from computational complexity that limits their universal usefulness[9]. Still some other video quality metrics, such as SSIM and multiscale MS-SSIM[10], have emerged that strike a good and useful balance between complexity and ability to predict human opinions with reasonable accuracy.

Another important class of video quality metrics analyzes primarily the signal characteristics of images, though they often also include some aspect of the human visual system. The VIF metric developed by Sheikh and Bovik[11], for example, incorporates the statistics of natural scenes[12]. Nill and Bouzas[13] developed an objective video quality metric based on the approximate invariance of the power spectra images. Lui & Laganiere[14,15] developed a method of using phase congruency to measure image similarity related to work by Kovesi[16,17] and based on the proposal by Morrone & Owens[18] and Morrone & Burr[19] and that perceptually significant features such as lines and edges are the features in an image where the spatial frequency components come into phase with each other. More recently, Zhang et al.[20] leveraged the concept of phase congruency to develop FSIM, a feature similarity metric.

The metric we propose in this paper falls in with the above group of metrics. It shares the same mind space in that it references statistically expectable spatial frequency statistics and the significance of phase information in an image; but also it differs in several important aspects. The metric we propose does not rely on phase congruency but rather on a “Spatial Detail” signal that can be thought of as a combination of the true phase information in an image and the statistically unpredictable information in any particular image. The “Spatial Detail” signal can be thought of as the condensed essence of an image that has the twin advantages of being very easy to calculate and of providing a guide to the bright and dark features and textures that give HDR WCG its special appeal.

The Performance of Existing HDR Video Quality Metrics

It would be simple if we could use the SDR objective video quality metrics we have come to know so well to quantify HDR video quality also. It turns out that objective video quality assessment for HDR is not simple. HDR video quality assessment needs either new algorithms and metrics or a new more perceptually meaningful way of representing image data. Perhaps both will be needed.

Hanhart, et al.[1], recently reported a study of objective video quality metrics for HDR images. They looked at the accuracy, monotonicity, and consistency of a large number of both legacy SDR and newer HDR-specific metrics[21-24] with respect to each metric's prediction of subjective video quality scores. They found that metrics such as HDR-VDP-2[23] and HDR-VQM[24] that were designed specifically for HDR content were best.

Interestingly, Hanhart et al. also found that the performance of most full-reference metrics, including PSRN and SSIM, was improved when they were applied to nonlinear perceptually transformed luminance data (PU[25] and PQ[26]) instead of linear luminance data. A similar conclusion was reported earlier by Valenzise et al.[27] who used a perceptually uniform "PU transform" developed by Aydin et al.[25] to assess compressed HDR images. They found that PU-based PSNR and SSIM performed as well and sometimes better than the more computationally demanding HDR-VDP[21] algorithm. Another study by Mantel et al.[28] also reported that perceptual linearization influenced the performance of objective metrics, though in this study perceptual linearization did not always improve performance. Rerabek et al.[29] extended the study of objective metrics beyond still images to HDR video sequences and found that perceptually weighted variants of PSNR, SSIM, MSE, and VIF correlated well with subjective scores, though HDR-VDP-2 was found to be the best performer statistically.

Balancing Performance and Complexity

Objective video quality algorithms should be as simple as possible and no simpler. Complex models of human vision are important and have their place, but can also become too cumbersome to be practically deployed in production and distribution of video programs. On the other hand, simpler fidelity metrics such as PSNR, SSIM, and MS-SSIM might be setting the bar too low even with perceptually linearized image data.

This paper proposes new HDR WCG video distortion metrics and an algorithm that is intended to be simple, fast, and provide actionable data to monitor and improve everyday video operations.

The video distortion assessment method we present leverages a framework of biologically inspired image and video processing developed by McCarthy & Owen[30,31] based on studies of the vertebrate retina and the expectable statistics of natural scenes. This bio-inspired framework has been leveraged previously to develop a perceptual pre-processor used in profession broadcast encoders[32] to make video more compressible while minimizing introduced artifacts. The details of the theory are beyond the scope of the paper, but the applicable elements of the theory can perhaps best be explained by considering video in terms of spatial frequency (see Figure 2).

CHARACTERISTICS OF HDR WCG VIDEO

Test Sequences & Preparation

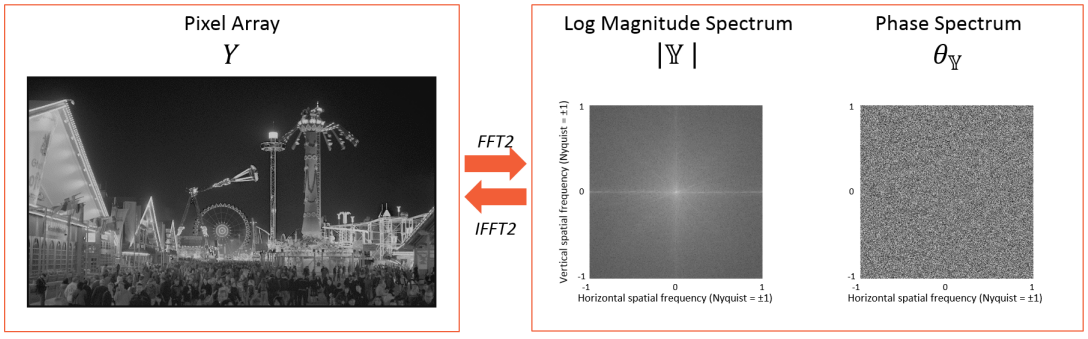
In this study, we used the HDR WCG test sequences shown in Figure 1. These sequences were created by the "HdM-HDR-2014 Project"[33,34] to provide professional quality cinematic wide gamut HDR video for the evaluation of tone mapping operators and HDR displays. All clips are 1920x1080p24 and color graded for Rec.2020 primaries & 0.005-4000 cd/m^2 luminance. To simulate cable and pay TV scenarios, we converted the original color graded frames (RGB 48 bits per pixel TIFF files) to YCbCr v210 format (4:2:2 10 bit) using the equations defined in ITU-R BT.2020[35]. All video processing and analysis was performed using Matlab[36], ffmpeg[37], and x265[38].

Bistro carousel_fireworks cars_fullshot hdr_testimage smith_hammering

Figure 1 - HDR WCG Test Sequences Used in this Study

Representing Images in Terms of Spatial Frequency

An image is normally thought of as a 2-dimensional array of pixels with each pixel being represented by red, green, and blue values (RGB) or a luma and 2 chroma channels (for example, YUV, YCbCr, and more recently IC$_T$C$_P$). An image can also be represented as a 2-dimensional array of spatial-frequency components as illustrated in Figure 2. The visual pixel-based image and the spatial-frequency representation of the visual image are interchangeable mathematically. They have identical information, just organized differently.



$$FFT2\big(Y(x,y)\big) = \mathbb{Y}(k_x,k_y) = \big|\mathbb{Y}(k_x,k_y)\big| * exp\left(i\theta_{\mathbb{Y}}(k_x,k_y)\right)$$

Figure 2 - Representation of a Video Frame in Terms of Spatial Frequency

Spatial-frequency data can be obtained from an image pixel array by performing a 2-dimensional Fast Fourier Transform (FFT2). The pixel array can be recovered by performing a 2-dimensional Inverse Fast Fourier Transform (IFFT2). FFT2 and IFFT2 are well known signal processing operations that can be calculated quickly in modern processors.

In the spatial frequency domain, the information in an image is represented as a 2-dimensional array complex numbers; or equivalently as the combination of a real-valued 2-d magnitude spectrum and a real-valued 2-d phase spectrum. (Note that the log of the magnitude spectrum is shown in Figure 2 to aid visualization. The horizontal and vertical frequency axes are shown relative to the corresponding Nyquist frequency (±1).)

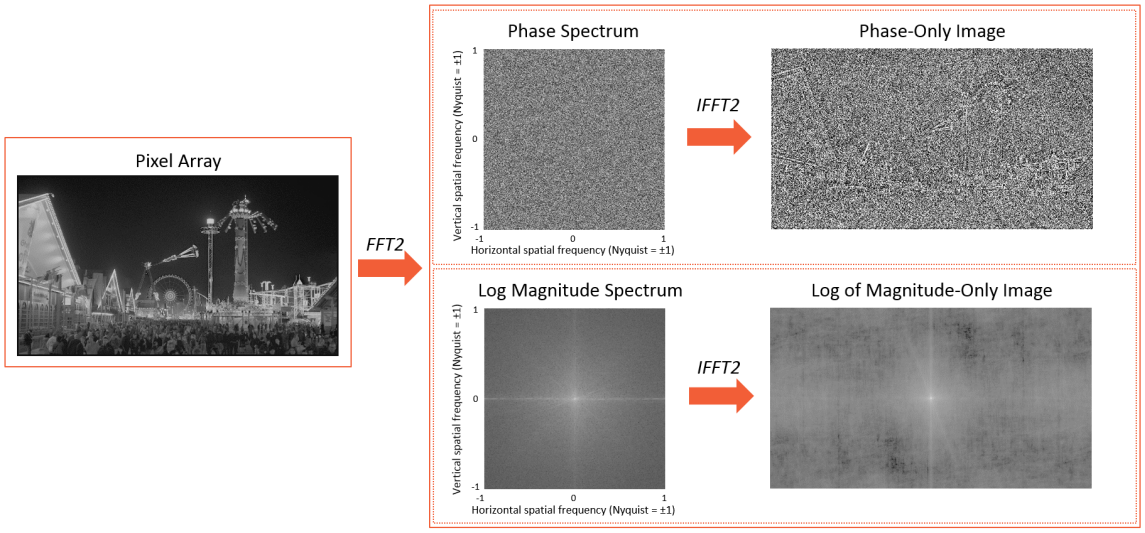


Figure 3 - The Phase Spectrum Typically Contains Most of the Details of an Image

The phase spectrum contains most of the specific details on the image, as illustrated in Figure 3. One way to think of the phase spectrum is that it provides information on how the various spatial frequencies interact to create the features and details we recognize in images[18,19]. The magnitude spectrum typically carries little unique identifying information about an image. Instead, it provides information on how much of the overall variation within the visual (pixel-based) image can be attributed to a particular spatial frequency.

Expectable Statistics of Complex Images

Images of natural scenes have an interesting statistical property: They have spatial-frequency magnitude spectra that tend to fall off with increasing spatial frequency in proportion to the inverse of spatial frequency[12]. The magnitude spectra of individual images can vary significantly; but as an ensemble-average statistical expectation, it can be said that “the magnitude spectra of images of natural scenes fall off as one-over-

spatial-frequency." This statement applies to both horizontal and vertical spatial frequencies.

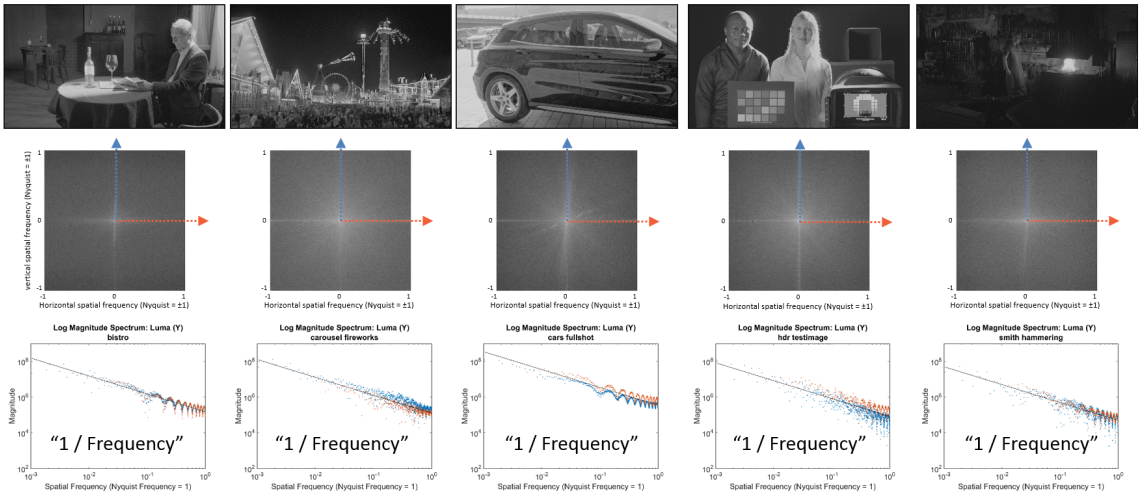


Figure 4 - Illustration of "One-Over-Spatial-Frequency" Magnitude Spectra

Figure 4 demonstrates that individual frames of the HDR WCG test sequences used in this study generally adhere to the "one-over-spatial-frequency" statistical expectation. The plots along the bottom of the figure show the values of the magnitude spectrum along the principal horizontal (orange) and vertical (blue) axes corresponding to the horizontal (orange) and vertical (blue) arrows in the middle row of the figure.

It is worth noting that the expectable statistics of "natural-scene" images are not limited to pictures of grass and trees and the like. Any visually complex image of a 3-dimensional environment tends to have the one-over-frequency characteristic, though man-made environments tend to have stronger vertical and horizontal bias than unaltered landscape. The one-over-frequency characteristic can also be thought of as a signature of scale-invariance, which refers to the way in which small image details and large image details are distributed. Images of text and simple graphics do not tend to have one-over-frequency magnitude spectra.

PROPOSED HDR WCG VIDEO DISTORTION ALGORITHM

Spatial Detail

HDR is all about preserving spatial detail. It is not about brighter pictures[39,40], or at least it should not be. The wider luminance range encoded by HDR enables crisp spatial detail

in dark regions and bright highlights to play a role in storytelling that is not possible otherwise. Similarly, WCG is all about enabling colorfulness of spatial details.

What is “spatial detail?” We know it when we see it; but if we can’t measure it quantitatively we can’t manage it systematically.

We propose that “spatial detail” can be quantified as the phase information in an image combined with the statistically unexpectable variations in the magnitude spectrum information.

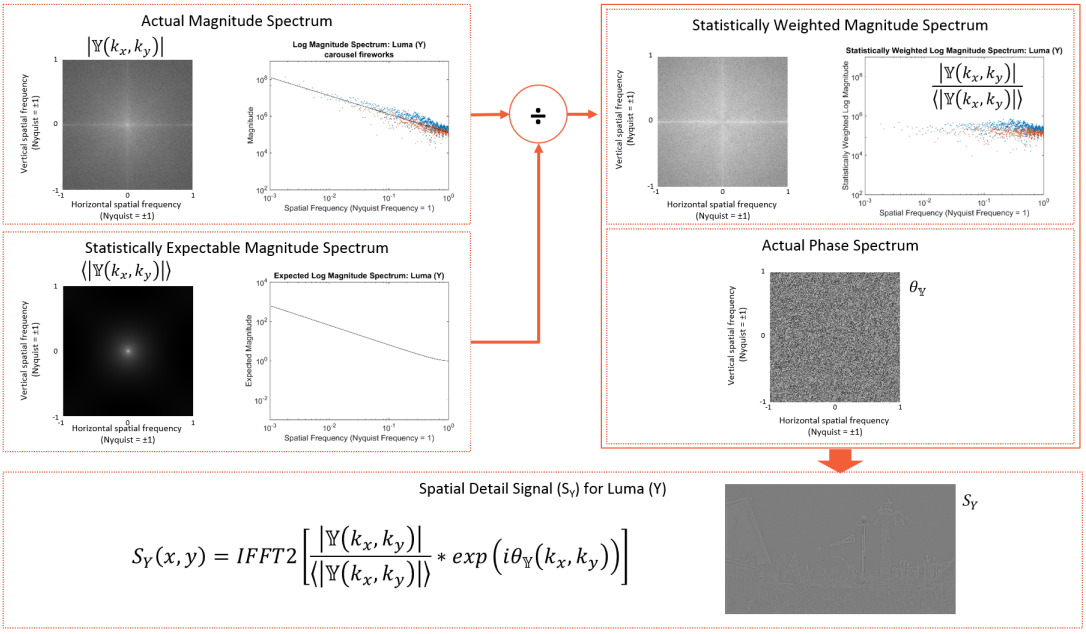


Figure 5 - Method of Calculating the Spatial Detail Signal

Our method of creating a Spatial Detail signal is illustrated in Figure 5. First, the magnitude and phase spectra are calculated from the image pixel array (only the luma channel is shown in Figure 5, but the methodology may also be applied to the chroma channel or, alternatively, to the red, green, and blue channels.) Next, a predetermined archetype of the statistically expectable one-over-frequency magnitude spectrum is divided into the actual magnitude spectrum to produce a statistically weighted magnitude spectrum. Third, the statistically weighted magnitude spectrum is combined with the actual phase spectrum. Finally, a 2-dimensional Inverse Fast Fourier Transform is applied to produce a pixel array that we call the Spatial Detail signal (see Figure 6).

Figure 6 - Enlarged View of the Spatial Detail Signal for the Luma Component

The Spatial Detail signal can be thought of as the result of a “whitening” process. However, a true whitening is a signal processing operation that results in exactly equal magnitude values at all frequencies. The phase image shown in Figure 3 is the result of a true whitening process. It is perhaps more useful and accurate to think of the Spatial Detail as the result of “statistically expectable whitening” that contains the result of a true whitening (the phase image) filtered by the statistically unexpectable modulations of the magnitude spectrum. The distinction might seem nuanced, yet the difference has practical benefits. Whereas the phase image (Figure 3) is rough and “noisy” in a way that obscures the recognizable details in an image, the Spatial Detail signal (Figure 6) is a smoothly varying more recognizable dual of the original image.

The Spatial Detail signal may also be thought of as the result of a true 2-dimensional differentiation of the image pixel array. The Spatial Detail signal is obtained by dividing the actual magnitude spectrum by a one-over-frequency spectrum, which is equivalent to multiplying the actual magnitude spectrum by frequency. Multiplication by frequency in the frequency domain is equivalent to differentiation in the pixel domain.

The differentiation characteristic of the Spatial Detail is apparent in Figure 7. The luma values of the original pixel array (***A***) along the midline (dashed line) are plotted in the upper middle graph (***C***). The histogram of the all the luma values of the original pixel array are plotted in the upper right graph (***E***). The corresponding Spatial Detail signal (***B***) values along the midline are plotted in the lower middle graph (***D***). The histogram of the all the Spatial Detail values are plotted in the lower right graph (***F***). Note that the Spatial

Detail values tend to cluster near zero and deviate significantly from the zero line only where the original luma values change significantly.

Note also that the Spatial Detail histogram is centered on zero and is symmetric, biphasic, and forms a compact peaked distribution. Conversely, the original luma values are spread out. The significance of this distinction is that the distribution of Spatial Detail values is preserved across images. The width of the histogram changes moderately from one video sequence to another but retains the stereotypical compact, peaked, biphasic, and symmetric shape. In other words, the Spatial Detail distribution is statistically expectable in the same sense that the one-over-frequency magnitude spectrum is statistically expectable. The histogram of original luma values is not statistically expectable: It changes significantly from one video sequence to another and even between scenes of the same program.

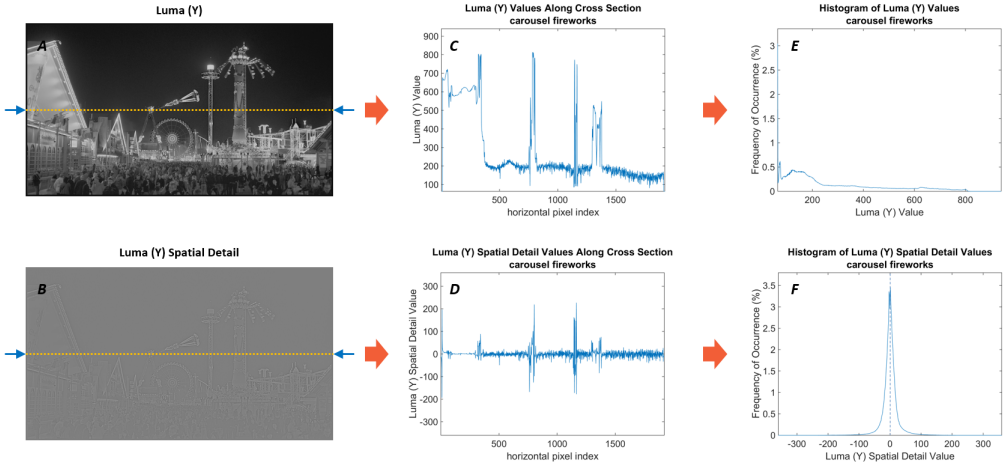


Figure 7 - The Spatial Detail Signal Distribution is Compact, Symmetric, & Biphasic

Effect of HEVC Compression on Spatial Detail Correlation

The Spatial Detail signal might be thought of as the condensed essence of the original image. As such, we explored the possibility that changes in the Spatial Detail signal that result from compression might prove to be a useful indicator of distortions and artifacts.

We used a 10-bit build of x265 (HEVC) to compress each of the test sequences at five different levels using the "constant quality" *crf* parameter (10, 15, 20, 25, and 30). The input to x265 in each case was the YCbCr 4:2:2 10-bit version of the original content. The internal x265 compressed pixel format was set as YCbCr 4:2:0 10-bit to simulate cable & pay TV workflows. The resulting average bitrates are plotted in Figure 8. We then decoded each frame of each compressed bitstream to YCbCr 4:2:2 10-bit for direct comparison with the input.

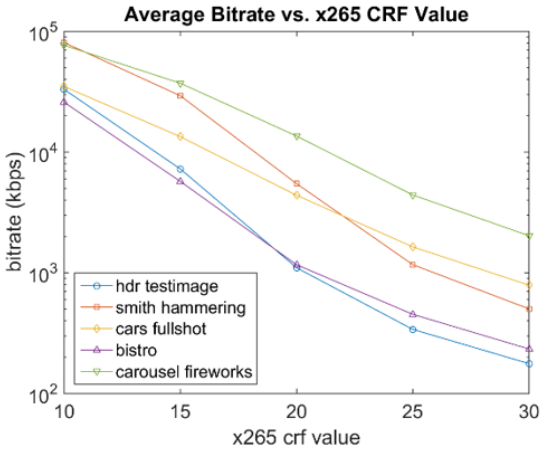


Figure 8 - Bitrates for HEVC-Compressed Test Sequences

We discovered that simple correlation analysis of the Spatial Detail signals provides a useful metric. The correlation of the luma values of the uncompressed pixel arrays (horizontal axis) and corresponding compressed pixel array (vertical axis) are shown in the upper row of Figure 9 for *crf* values 10 (middle) and 30 (right). The analogous graphs on the lower row are for the values of the corresponding Spatial Detail signals. If the uncompressed and compressed values were identical the data points would describe a perfect line of unity slope. Differences between the uncompressed and compressed data cause a scatter about the line. More compressed data (larger *crf* value) can be expected to result in a larger amount of scatter. Note though that the change in scattering is more pronounced for the Spatial Detail signal than the original luma values. More compression causes the scatter of the Spatial Detail values to become more globular, becoming more compact along the line of perfect correlation and expanding perpendicular to that line.

The amount of scatter – the amount of uncorrelation – is quantifiable by the coefficient of determination, R^2 (pronounced "R-squared"), which is a statistical measure of the amount of predictability of one data set given another data set. In our case of simple linear regression, R^2 is simply the square of the Pearson correlation coefficient. An R^2 value of 1 means perfectly correlated and a value of 0 means perfectly uncorrelated.

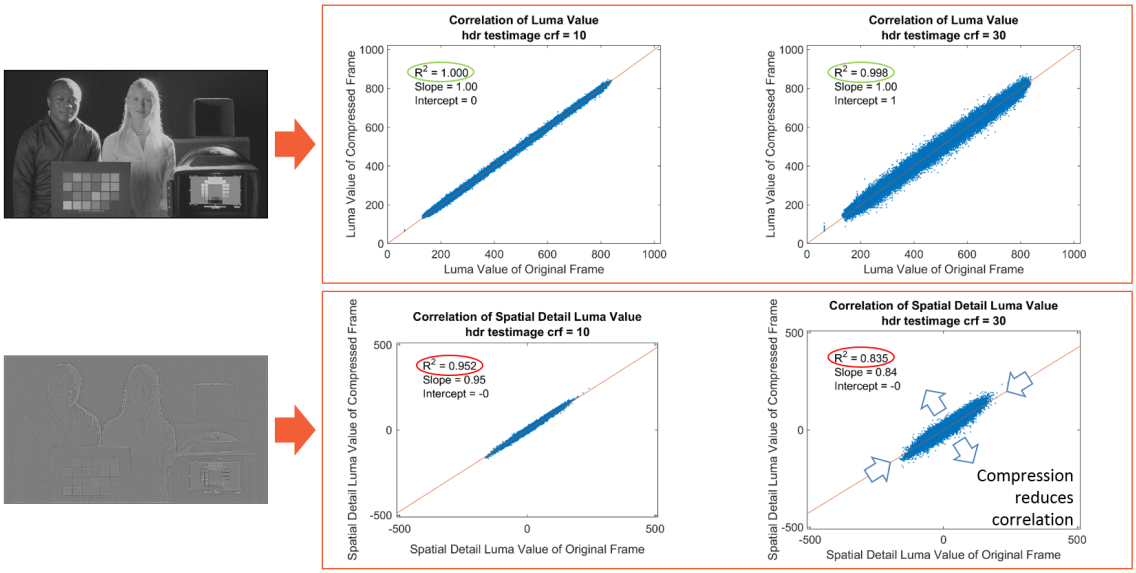


Figure 9 - Correlation of Luma and Corresponding Spatial Detail Signals

R^2 values for all the test sequences at every compression level are plotted in Figure 10. For the original luma values (right-hand graph), the value of R^2 changes only slightly between *crf* values of 10 and 30 even though the bitrate changes by approximately 2 orders of magnitude (see Figure 8). For the corresponding Spatial Detail signal, the story is very different (left-hand graph). The value of R^2 changes significantly over the same range of *crf* values and corresponding bitrates.

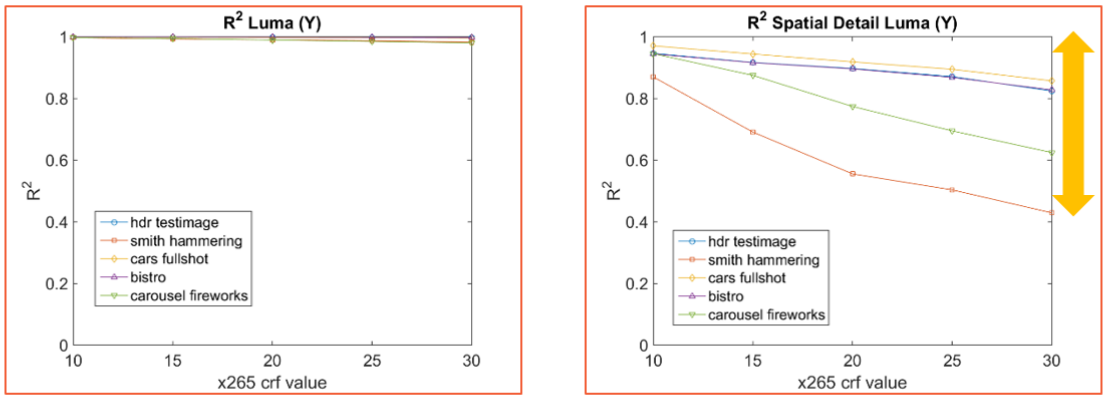


Figure 10 - Correlation Values for All Test Sequences & HEVC Compression Levels

Results from well-established video quality metrics for the same test sequences and compression levels are plotted in Figure 11 to provide a point of comparison and reference. PSNR displays good sensitivity over the entire range. MS-SSIM is also sensitive to compression in the range that can be expected in cable and pay TV service, but only over a very tiny restricted range of values from 0.98 to 1 out of a full range of 0

to 1. In comparison, R^2 values for Spatial Detail ranges from 0.4 to 1 out of a full range of 0 to 1.

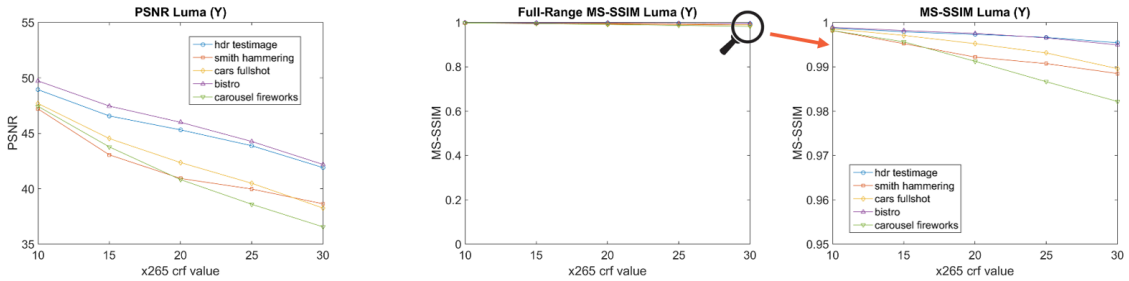


Figure 11 - PSNR and MS-SSIM Values for All Test Sequences & Compression Levels

Using Spatial Detail to Probe Bright & Dark Features and Textures

The Spatial Detail signal can be decomposed into two subcomponents (Figure 12) that can be used as guides for selectively analyzing perceptually significant features and textures. A "Sign" map (lower left in Figure 12) of the Spatial Detail signal can be created simply as a binary image in which each pixel is set to 0 if the corresponding Spatial Detail pixel is negative and set to 1 if it is positive. The Sign map will tend to have an equal number of 0's and 1's because of the statistically expectable symmetric biphasic distribution of Spatial Detail values. A "Significance" map (lower right in Figure 12) can be created simply as the absolute value of the Spatial Detail signal. Bright regions of the Significance map correspond to larger absolute values of the Spatial Detail signal. Note that the Significance map tends to highlight edges, borders, and other transitions which is in-line with thinking of the Spatial Detail signal as a result of a true 2-dimensional spatial differentiation as discussed above.

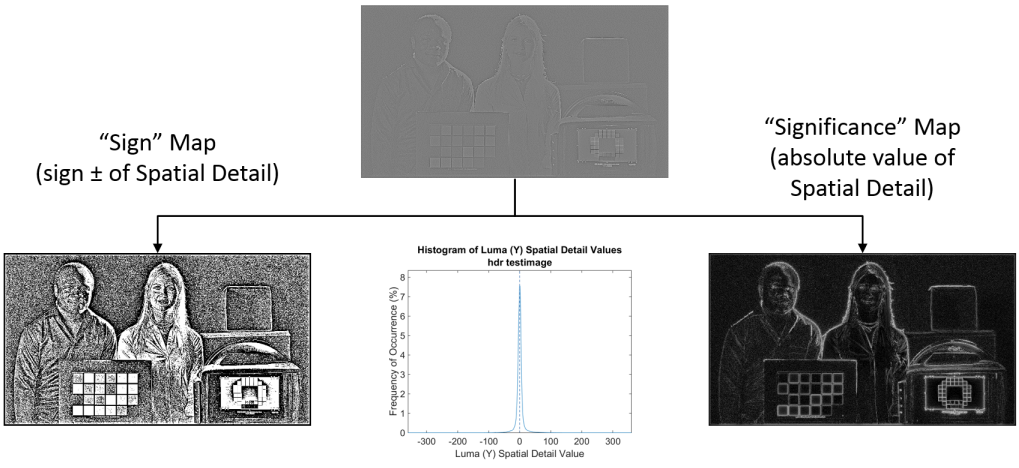


Figure 12 - Decomposition of Spatial Detail into a Sign map and a Significance map

The Spatial Detail signal can also be decomposed as illustrated in Figure 13 to provide a guide to “bright features”, “dark features”, and “textures”. Large positive values of the Spatial Detail signal can be used to define the location of bright features. Larger negative values can be similarly used to define the location of dark features. The remaining smaller positive and negative values of the Spatial Detail signal thus define textures. Absolute thresholds could be used but we find it more useful to use graded weighted functions such as but not limited to the following:

$$W_{bright}(x,y) = \frac{|S(x,y)|}{|S(x,y)| + S_0}(S(x,y) > 0)$$

$$W_{dark}(x,y) = \frac{|S(x,y)|}{|S(x,y)| + S_0}(S(x,y) < 0)$$

$$W_{texture}(x,y) = 1 - W_{bright}(x,y) - W_{dark}(x,y)$$

where $W_{bright}(x,y), W_{dark}(x,y),$ and $W_{texture}(x,y)$ are pixel array weighting maps having values between 0 and 1, and $S(x,y)$ is the Spatial Detail signal derived from the uncompressed luma component, and S_0 is a tuning parameter that adjusts the boundary between feature and texture (equivalent to the vertical dashed lines in the top center graph of Figure 13).

The image in the middle of the lower row of Figure 13 was obtained by multiplying each red, green, and blue color plane by $W_{texture}(x,y)$. The lower right image illustrating the bright features was created the same way, but with $W_{bright}(x,y)$. The lower left was created using $W_{dark}(x,y) + W_{bright}(x,y)$ to visualize all features. (The weighting map in each case was calculated using the Spatial Detail signal of the luma component.)

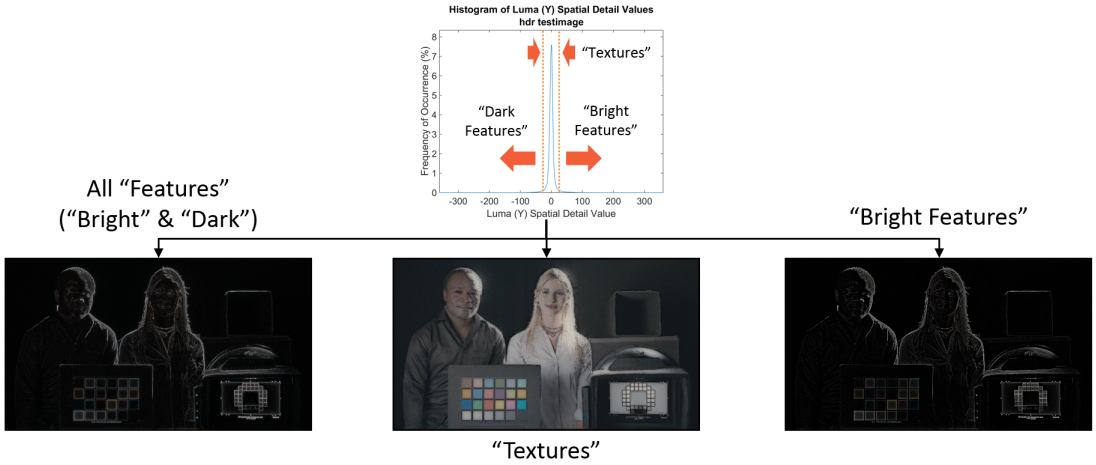


Figure 13 - Bright Features, Dark Features, and Textures

The proportion of the image that may be described as bright features, dark features, and textures may be quantified using formulae of the type below for an NxM sized video frame:

$$P_{bright} = \frac{\sum_{x,y}^{N,M} W_{bright}(x,y)}{NM}; P_{dark} = \frac{\sum_{x,y}^{N,M} W_{dark}(x,y)}{NM}; P_{texture} = \frac{\sum_{x,y}^{N,M} W_{texture}(x,y)}{NM}$$

Textures account for the majority of each of the HDR WCG test sequences though features play a relatively larger role in for “smith_hammering” and “carousel_fireworks”, as illustrated in Figure 14.

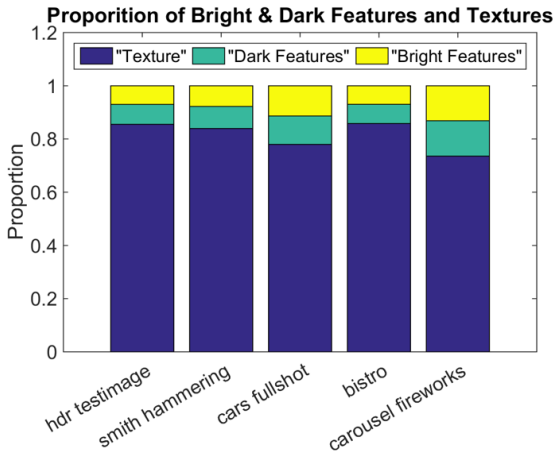


Figure 14 - Relative Proportions of Bright Features, Dark Features, and Textures

Spatial Detail Correlation for HDR WCG Features and Textures

Bright and dark features and textures are particularly important in HDR WCG video. They are what make HDR pop. We used correlation analysis to see if the bright features, dark features, or textures were systematically affected by HEVC compression preferentially.

The resulting R^2 values are plotted in Figure 15. We found that HEVC did a particularly good job of preserving both the bright and dark features even at compression levels beyond that which would normally be used in cable and pay TV services. Throughout the range of compression levels we tested, the R^2 values for all features remained above 0.9. The results for texture were not as good. R^2 values for texture dropped below 0.9 even for light HEVC compression thus indicating significant distortion. These findings were consistent across the test sequences thus indicating a systematic characteristic of HEVC compression rather than a content-dependent effect.

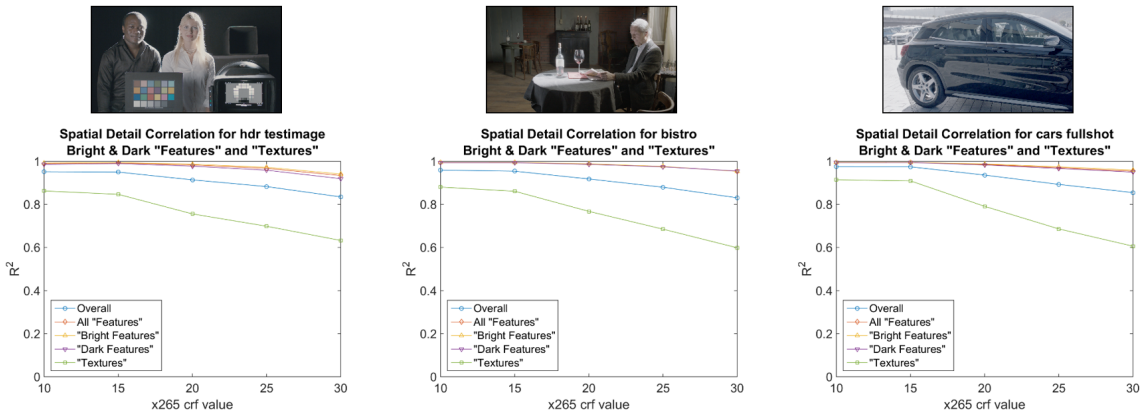


Figure 15 - Spatial Detail Correlation for Bright & Dark Features and Textures

Weighted Mean-Squared Error

We also investigated selective distortion for bright & dark features and textures using weighted Mean-Squared Error (MSE). The weighting was achieved by multiplying the squared difference between the uncompressed and compressed video frame data before summation over all pixels (frame size of NxM), as illustrated in the equations below.

$$MSE_{total} = \frac{\sum_{x,y}^{N,M} \left(Y_{ref}(x,y) - Y_{tst}(x,y)\right)^2}{NM}$$

$$MSE_{total} = MSE_{bright} + MSE_{dark} + MSE_{texture}$$

$$MSE_{bright} = \frac{\sum_{x,y}^{N,M} W_{bright}(x,y) \left(Y_{ref}(x,y) - Y_{tst}(x,y)\right)^2}{NM}$$

The values of MSE_{dark} and $MSE_{texture}$ may be calculated in a similar manner. The resulting weighted MSE values provide insight into the proportion of the total MSE may be attributed to bright & dark features and textures. The same methodology may be applied to both luma and chroma MSEs with appropriate scaling for the 4:2:2 YCbCr format.

Weighted MSE results for the HDR WCG test sequences that are plotted in Figure 16 demonstrate that the majority of the total MSE is attributable to the texture component. We found this conclusion to be consistent across all HDR WCG test sequences for all compression levels we tested and that the conclusion holds for luma and chroma. The dominance of texture MSE is mainly a result of texture making up the largest proportion of video frames (see Figure 14).

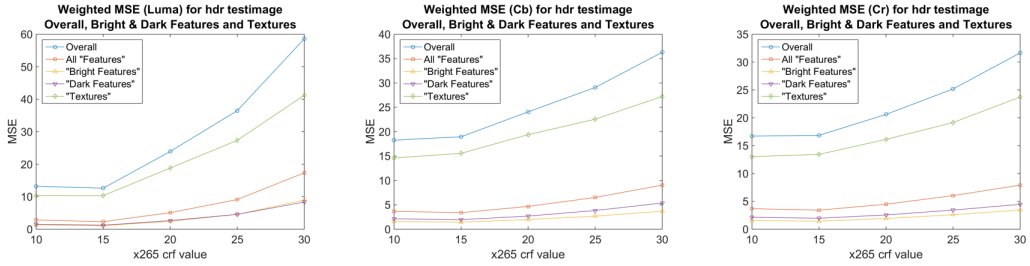


Figure 16 - Weighted MSE for Bright & Dark Features and Textures

Squared-Error Density

Introduction of a Squared-Error Density (SED) provides a means of selectively probing distortion for features and textures while accounting for each one's relative prominence in HDR WCG video. SED may be calculated for bright & dark features, and textures according to the following equations:

$$SED_{bright} = \frac{MSE_{bright}}{P_{bright}}; SED_{dark} = \frac{MSE_{dark}}{P_{dark}}; SED_{texture} = \frac{MSE_{texture}}{P_{texture}}$$

SED is MSE divided by the corresponding proportionality of feature or texture. SED thus accounts for the fact that features tend to be rarer the texture (see Figure 14). SED may be thought of as providing a measure of equitability between features and textures. For example, SED can provide insight into whether rarer features experience disproportionate distortion compared to texture.

SED results for the HDR WCG test sequences are plotted in Figure 17. We find squared-error density for bright and dark features is relatively more severe than for textures. This finding is consistent for all HDR WCG test sequences and compression levels we tested and holds for luma and chroma.

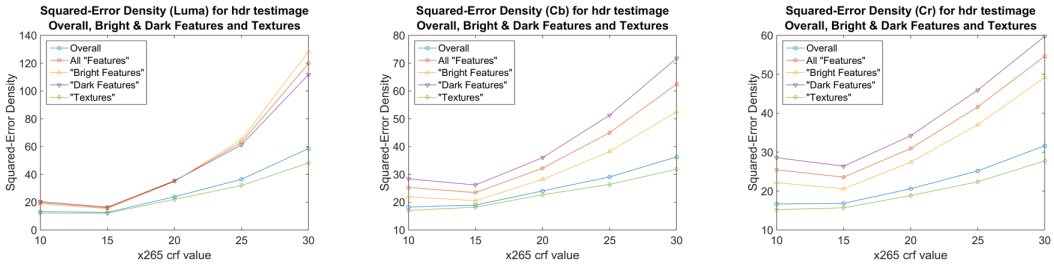


Figure 17 - Squared-Error Density for Bright & Dark Features and Textures

CONCLUSION

We have presented in this paper a set of video distortions metrics that might prove to be particularly useful for HDR WCG video. The main motivating principle we presented was the “Spatial Detail” signal that we used in two ways: 1) as a proxy for the original image data; and 2) as a guide to the perceptually important “features” and “textures” in HDR WCG video.

The Spatial Detail signal is a condensed version of the original image that preserves the recognizable details in an image while discounting local luminance. It can be thought of as a true 2-dimensional differential of the original image. It may also be understood in terms of the phase information in an image in conjunction with the statistically unpredictable information in an image. From a practical standpoint, it doesn’t really matter which theory one prefers. Instead, an important key characteristic of the Spatial Detail signal is that it has a statistically stable and expectable compact, peaked, biphasic, and symmetric distribution of values that is preserved across a wide range in images and video. Larger values – positive and negative – form a convenient guide to the kinds of features people tend to find significant. Spatial Detail values nearer the zero midpoint of the distribution form a convenient guide to image regions that people would tend to classify as textural. Such feature and texture maps provide a stable framework in which to selectively investigate the perceptual potent highlights and dark details that are the hallmark of HDR WCG video.

We presented three HDR WCG video distortion metrics in this paper:

1. For the first metric, we used Spatial Detail as a proxy for the original image and showed that correlation between the Spatial Detail signals of the uncompressed and compressed versions of HDR WCG video was systematically affected by the aggressiveness of HEVC compression. By combining Spatial Detail correlation with our feature and texture assignment methods, we showed that texture correlation was impacted significantly more than feature correlation.
   Spatial Detail correlation has several distinctions when compared to established video quality metrics. It can be used selectively on bright & dark features and on textures. Moreover, Spatial Detail values are in the range of 0 to 1, which is more intuitive than the unbounded PSNR scale, while being a much more sensitive indicator than MS-SSIM over the range of compression levels typical of cable and pay TV operations.

2. For the second metric, we used Spatial Detail as a guide for bright & dark features and texture to selectively quantify the MSE for each layer of image detail. We showed that texture is the largest contributor to overall MSE mainly, because texture regions typically make up a larger proportion of any image than the rarer feature regions.

3. The third metric introduced a Squared-Error Density (SED) that compensates for the relative proportions of feature and texture in an image so as assess distortions on a more equal scale. We found that SED indicates that features experience disproportionate distortion compared to texture.

We have deliberately used the term “video distortion” instead of “video quality” throughout this paper. The main reason for doing so is that the metrics we proposed have not yet been compared to subjective test scores and thus may not yet be claimed to be calibrated subjective quality metrics. Also, it is not the intent of this paper to link the metrics we propose to subjective assessment; though we may do so in latter publications. Rather, our intent is to provide easy to calculate metrics that we hope can provide insight during this critical period in our industry as we work through the technical and creative issues related to HDR and WCG.

It is also worth highlighting that the Spatial Detail signal and related metrics are easy to calculate using modern signal processing techniques in modern processors. Thus, we believe the technical barrier to adoption of these metrics is low.

Our intent in the paper is to provide useful and easy-to-calculate metrics that have a low technical barrier to adoption. The Spatial Detail signal and related metrics we propose are easy enough to calculate that they are candidates for real-time HDR WCG video assessment using modern signal processing techniques in modern processors. Our next steps will be to continue to assess the utility of our HDR WCG metrics with the hope that they will help MSOs navigate key technical and creative issues as HDR WCG video programming emerges as the next wave of great subscriber experiences.

ABBREVIATIONS

FFT2	2-dimensional Fast Fourier Transform
FSIM	Feature-Similarity Index
HDR	High Dynamic Range
HEVC	High Efficiency Video Coding
ICTCP	ICTCP color space
IFFT2	Inverse 2-dimension Fast Fourier Transform
MSE	Mean Square Error
MSO	Multiple Systems Operators
MS-SSIM	Multiscale Structural Similarity
PQ	Perceptual Quantizer
PSNR	Peak Signal-to-Noise Ratio
PU	Perceptually Uniform
SDR	Standard Dynamic Range
SED	Squared-Error Density
SSIM	Structural Similarity
YCbCr	YCbCr color space
VDP	Visual Difference Predictor
VIF	Visual Information Fidelity
VQM	Video Quality Measure
YUV	YUV color space
WCG	Wide Color Gamut

RELATED READINGS

- **A Systematic Approach to Video Quality Assessment and Bitrate Planning** – In this paper, the author presents a streamlined method of setting operational video quality and bandwidth using either subjective or objective testing, using individual golden-eyes or focus groups of any size. The data and analysis included are intended to aid in planning video quality and bandwidth resources across a range of service offerings from OTT through Ultra HD.
- **Efficient Content Processing for Adaptive Video Delivery** – This paper provides an in-depth overview of two emerging technologies, dynamic profile selection and cooperative transcoding, along with experimental data demonstrating their potential for substantially reducing content processing requirements for multiscreen video delivery.
- **Methodologies for QoE Monitoring of IP Video Services** – This paper discusses the differences between QoE and QoS and between QoE and video quality and then compares different methodologies for video quality and QoE monitoring. It also includes a review of alternatives for embedding QoE probes in the end-to-end IP Video architecture and their ability to collect true and effective QoE information.

MEET OUR EXPERT: Sean T. McCarthy

Dr. Sean McCarthy, Fellow of the Technical Staff, brings a unique convergence of expertise in video compression, signal processing, and the neurobiology of human vision to content distribution at ARRIS. Dr. McCarthy leads advancements in state-of-the-art of video processing, compression and practical vision science. Previously, he held similar responsibilities as Fellow of the Technical Staff at Motorola, and as Chief Scientist at both Modulus Video, which was acquired by Motorola. Prior to that, Dr. McCarthy had similar responsibilities at ViaSense, a University of California, Berkeley spin-off that developed commercial applications of the human visual system. He earned a B.S. in physics from Rensselaer Polytechnic, and earned his Ph.D. in bioengineering jointly at University of California, Berkeley and University of California, San Francisco.

REFERENCES

(1) Hanhart, P., Kroshunov, P., and Ebrahimi, T. “Subjective evaluation of higher dynamic range video.” Proceedings of SPIE - The International Society for Optical Engineering, 2014

(2) Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004

(3) Hanhart, P., Bernardo, M.V., Korshunov, P., and Pereira, M. “HDR image compression: A new challenge for objective quality metrics.” in Sixth International Workshop on Quality of Multimedia Experience (QoMEX), 2014

(4) ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, Jan. 2012

(5) ITU-T P.910, “Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union, April 2008

(6) Winkler, S. Digital Video Quality: Vision Models and Metrics, John Wiley & Sons, Mar. 2005

(7) VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” Mar. 2000. http://www.vqeg.org/.

(8) Wang, Z. and Bovik, A. C. “Mean squared error: love it or leave it? - A new look at signal fidelity measures,” IEEE Signal Processing Magazine, vol. 26, no. 1, pp. 98-117, Jan. 2009

(9) Hanhart, P., Bernado, M.V, Pereira, M., Pinheiro, M.G., and Ebrahimi, T. “Benchmarking of objective quality metrics for HDR image quality assessment.” EURASIP J. Image & Video Processing, 2015

(10) Wang, Z., Simoncelli, E.P., and Bovik, A. “Multi-scale structural similarity for image quality assessment.” Proc. of the 37th IEEE Asilomar Conference on Signals, Systems, and Computers, 2003

(11) Sheikh, H. R. and Bovik, A. C. “Image information and visual quality,” IEEE Transactions on Image Processing, vol. 15, no. 2, pp. 430–444, Feb. 2006

(12) Field, D.J. “Relationship between the statistics of natural images and the response properties of cortical cells.” J. Opt. Soc. Am. A. Vol. 4, No. 12, 1987

(13) Nill, N. B. and Bouzas, B. H. “Objective image quality measure derived from digital image power spectra.” Optical Engineering, vol 31, no. 4, 1992

(14) Liu, r. and Laganiere, R. “On the use of phase congruency to evaluate image similarity.” IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006

(15) Liu, r. and Laganiere, R. “Phase congruence measurement for image similarity assessment.” Pattern Recognition Letters, vol 28, no. 1, 2007

(16) Kovesi, P. “Image Features from Phase Congruency.” in Videre: Journal of Computer Vision Research, Vol 1, No. 3, The MIT Press, 1999

(17) Kovesi, P. “Invariant measures of image features from phase information.” Thesis (Ph.D.) Dept. of Computer Science. University of Western Australia, 1996

(18) Morrone, M.C. and Owens, R.A. “Feature detection from local energy.” Pattern Recognition.” Lett., 303–313, 1987

(19) Morrone, M. C. and Burr, D. C. “Feature detection in human vision: A phase-dependent energy model.” Proc. Royal Soc. Of London, Series B, Biological Sciences, vol. 235, no. 1280, 1988

(20) Zhang, L., Zhang, L., Mou, X., and Zhang, D. “FSIM: A feature similarity index for image quality assessment.” IEEE Trans. Image Process. vol 20, no.8, 2011

(21) Mantiuk, R., Daly, S., Myszkowski, K., and Seidel, H.-P. “Predicting visible differences in highdynamic range images: model and its calibration.” SPIE Human Vision and Electronic Imaging X, vol. 5666., 2005

(22) Daly, S.J. “Visible differences predictor: an algorithm for the assessment of image fidelity” SPIE Human Vision, Visual Processing, and Digital Display III, vol.1666., 1992

(23) Mantiuk, R., Kim, K.J., Rempel, A.G., and Heidrich, W. “HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions.” ACM Trans. Graph. 30(4), 40:1–40:14, 2011

(24) Narwaria, M., Mantiuk, R.K., Perreira Da Silva, M., and Le Callet, P. “HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images.” J. Electron. Imaging. 24(1), 010501, 2015

(25) Aydin, T. O., Mantiuk, R., and Seidel H.-P. “Extending quality metrics to full luminance range images.” Human Vision and Electronic Imaging XIII. Edited by Rogowitz, Bernice E.; Pappas, Thrasyvoulos N. Proceedings of the SPIE, Volume 6806, 2008

(26) Miller, S., Nezamabadi, M., and Daly, S., “Perceptual Signal Coding for More Efficient Usage of Bit Codes.” SMPTE Motion Imaging Journal, 2013

(27) Valenzise, G., De Simone, F., Lauga, P., and Dufaux, F. “Performance evaluation of objective quality metrics for HDR image compression.” Proc. SPIE 9217, Applications of Digital Image Processing XXXVII, 2014

(28) Mantel, C., Ferchiu, S. C., Forchhammer, S. “Comparing subjective and objective quality assessment of HDR images compressed with JPEG-XT.” in 16th International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2014

(29) Rerabek, M., Hanhart, P., Korshunov, P., and Ebrahimi, T. “Subjective and objective evaluation of HDR compression.” International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM, Chandler, Arizona, USA. February 2015

(30) McCarthy, S.T. and Owen, W.G. “Apparatus and Methods for Image and Signal Processing”. US Pat. 6014468 (2000). US Pat. 6360021 (2002), US Pat. 7046852 (2006), 1998

(31) McCarthy, S., “A Biological Framework for Perceptual Video Processing and Compression,” SMPTE Mot. Imag. J., 119(8):24-32, Nov/Dec. 2012

(32) McCarthy, S.T. “Theory and practice of perceptual video processing in broadcast encoders for cable, IPTV, satellite, and internet distribution.” Proc. SPIE 9014, Human Vision and Electronic Imaging XIX, 2014

(33) Froehlich, J., et al. “HdM-HDR-2014 Project,” http://www.hdm-stuttgart.de/~froehlichj/hdm-hdr-2014

(34) Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schillin, A., and Brendel, H. “Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays.” Proc. SPIE 9023, Digital Photography X, 2014

(35) ITU-R BT.2020 “Parameter values for ultra-high definition television systems for production and international programme exchange." International Telecommunication Union, 2012

(36) The Mathworks. www.mathworks.com

(37) ffmpeg. www.ffmpeg.org

(38) x265. www.x265.org

(39) ITU-R BT.2100. “Image parameter values for high dynamic range television for use in production and international programme exchange.” 2016

(40) Report ITU-R BT.2390-0 “High dynamic range television production and international programme exchange.” International Telecommunication Union, 2016