IMAGE ADAPTATION REQUIREMENTS FOR HIGH DYNAMIC RANGE VIDEO UNDER REFERENCE AND NON-REFERENCE VIEWING CONDITIONS

M. Pindoria and S. Thompson

British Broadcasting Corporation, United Kingdom

ABSTRACT

Until now, most of the research in the field of high dynamic range (HDR) video has centred on the use of non-real-time graded images which have been adjusted to look correct on a known reference screen in a reference environment. For live television, without the luxury of grading, it is important that images captured directly by the camera look correct. So the HDR system's end-to-end opto-optic transfer function (OOTF), which maps the light captured at the camera sensor to the light output from the display, is of paramount importance. Furthermore, it is critical that the artistic intent of the video is preserved when rendered for the viewer with a different screen in a different viewing environment.

The authors present results of two subjective tests. The first test determines the most suitable OOTF for a reference environment and display; the second test determines how this transfer function could be adjusted so the high dynamic range video signal can be displayed on a range of different brightness displays whilst maintaining artistic intent.

INTRODUCTION

High Dynamic Range video (HDR) is a relatively new technique which allows the content producer to more accurately reproduce an image without the suppression of highlights usually associated with conventional video. Experiments show that there is a preference for high dynamic range video displayed on a high brightness monitor over a conventional television displaying standard dynamic range (SDR) video content (Hanhart et. al. (1)).

The television viewing experience has traditionally been defined in terms of a reference screen (EBU Tech 3320(2), ITU-R BT.1886(3))) being viewed in a reference environment (Teear (4)). At the time of writing HDR displays range in peak brightness from approximately 500 cd/m^2 to approximately 4000 cd/m^2. The authors expect brightness levels to increase as technology matures and new technologies reach the market. It is therefore critical that an end-to-end transfer function is chosen that allows artistic intent to be maintained on screens of differing peak brightness.

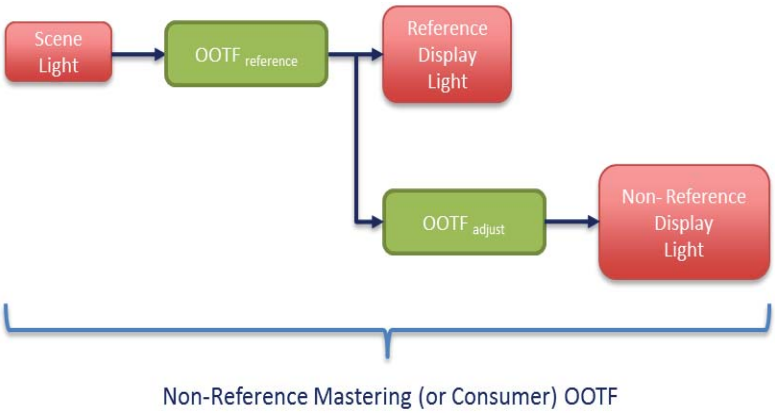

Figure 1 – Adjusting system gamma

In this document we first present results of experiments undertaken to find the most suitable reference OOTF. The reference OOTF is proposed for a reference HDR monitor, under reference viewing conditions. In the second set of experiments we explore this

reference OOTF further, to determine how it can be adjusted to display a high dynamic range video (Borer and Cotton (5)) signal on a range of different brightness screens under reference background lighting conditions. In these experiments we use an end-to-end power function, or system gamma as proposed in Borer and Cotton (5), as the OOTF and adapt it for different brightness displays by changing the value of gamma, as illustrated in Figure 1. Our results show the change in system gamma that is required to obtain the best perceptual match between a signal graded on a given screen in a reference environment, and the same signal on any other screen. Results of a supplementary test show that as the system gamma adjustments occur in the linear domain, they are independent of the exact choice of electro-optical or opto-electronic transfer functions.

BACKGROUND

The role of a reference OOTF is to map data from a scene captured by a camera to the display in a reference viewing environment. The reference OOTF should ideally be independent of any artistic adjustments so that, with adjustment to only the OOTF, the artistic intent can be maintained on a range of non-reference displays and in a range of viewing conditions. With legacy SDR, reference displays were similar to those TVs found in the home, predominately CRT displays, therefore adjustments to the OOTF were not deemed necessary. With HDR, and a wide range of display technologies and viewing brightnesses, the need for suitable adjustments to a reference OOTF is absolutely required to maintain the artistic intent of the video. Without any artistic adjustments or grading, such as in a live recording, the role of the OOTF is therefore to make the displayed images look as close as possible to the actual scene. Source linear scene-referred files were obtained from Fairchild (6) and from tests undertaken by Arnold & Richter (ARRI) and the Stuttgart Media University (Frölich et. al. (7)). The source images were not graded, the colours and tones within them represent the actual scene. Therefore, in the first experiment we asked our viewers to rate the images on “naturalness”, or “realism”. That is, “if they were standing next to the camera, how natural do the images look?” Note that this is distinct from asking the viewers to choose their preferred image, which would include an element of artistic appreciation. For this comparison we were looking to achieve the most realistic end-to-end system. The artistic choices to achieve the “best” image are a distinct and separate part of the television production process.

Without a reference image, rating the naturalness of an image is likely to be susceptible to user preference. Ashikhmin et. al. (8) carried out a series of subjective tests in which they were testing tone mapping operators. In their tests, they performed 3 experiments, in the first, they asked the viewer “which image do you like the most?”, and in the second test they asked “which image do you think is more real?”. For the initial 2 experiments, the results indicated that there was no clear distinction between the operators under test, and the results of each experiment were not well correlated. In the third experiment, the viewers were taken to a specific location where the image was taken, and then asked “which image is the closest to the real scene in front of you?” In this third experiment, the results were well correlated and there was a clear preference towards a particular tone mapping operator. The authors suggest that the concept of realism humans rely on, in the absence of a real scene (“abstract realism”), is rather imperfect and that it would be easier to identify consistently better performing algorithms if real scenes were used.

In our first experiment, as consistently reproducible scenes were unavailable in a lab environment for the duration of the tests, it was decided to use judgement based on naturalness to compare different OOTFs, noting that the standard deviation of the results will be high. The reference OOTFs under consideration were those submitted through the ITU-R Rapporteur’s Group on HDR (RG 24) (ITU-R (9)). A summary of the OOTFs is shown in Table 1. A full description of the OOTFs can be found in the ITU report.

Name	Summary
RGB_1.2	Overall system gamma of 1.2 applied on R, G, B components separately.
Y_1.4	Overall system gamma of 1.4 applied on luminance component only.
RGB_1.25	Overall system gamma of 1.25 applied on R, G, B components separately.
COMPOSITE SDR	Generalised OOTF from BT.1886 in combination with BT.709 applied on R, G, B components separately.

Table 1 – Reference OOTFs under test

In the second experiment, test subjects were asked to perceptually match as closely as possible an image displayed with a reference peak brightness to the same image with a different peak brightness by adjusting the system gamma applied to the non-reference image. Initial informal testing suggested that it was difficult to perceptually match screen brightnesses differing by a large ratio due to eye adaptation issues. The tests were therefore designed to use multiple intermediate reference screen brightnesses, (4000, 2000, 1000 cd/m^2) each only one stop apart.

Tests were undertaken using a SIM2 HDR47E display calibrated by BBC R&D using its colour calibrated logLUV input. An informal verification of the applicability of the results to non-LCD displays was undertaken using a prototype Sony BVM-X300 professional HDR OLED display.

METHODOLOGY

Test layout

A common test layout was used for both sets of subjective tests. The test room was set up so that the test subject could see the SIM2 and a monochromatic photographic background illuminated with a 5 cd/m^2 CIE D65 illuminant. A plan view is shown in Figure 2. We measured the ambient lighting to be 1.15 lux perpendicular to the screen with this lighting.

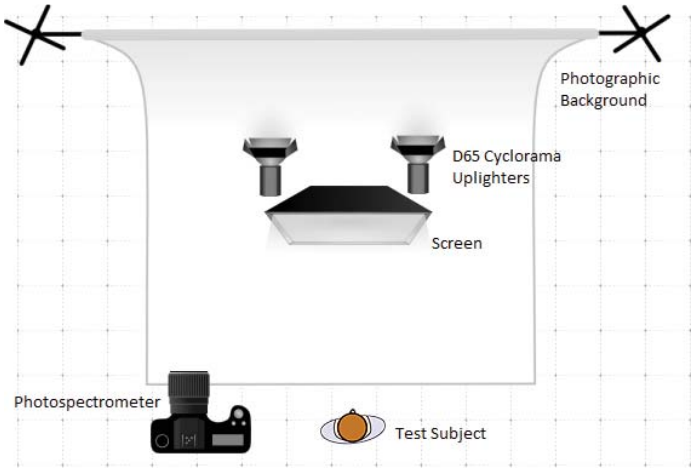

Figure 2 – Test Room Layout

The SIM2 display in HDR mode requires absolute luminance value logLUV images. Since the screen has no brightness control in its logLUV mode, we measured the required black level offset for the viewing conditions using a series of specially generated PLUGE signals, each with a different black level offset. An appropriate black offset was then added to the individual test images.

Comparison of reference OOTFs under reference viewing conditions

This subjective test was conducted using the ITU-R BT.500 test methodology (ITU-R (10)) using a Double Stimulus Continuous Quality Scale (DSCQS). Each of the OOTFs was compared against an anchor image and the viewers were asked to judge each image independently on the level of naturalness using a continuous scale, which was marked bad, poor, fair, good and excellent, which was translated to a scale from 0 to 100.

This test was conducted on a single screen, in an A, B, A, B type comparison, timing intervals are shown in Figure 3. Either A or B (or both) was a hidden anchor in every test. The presentation of the images and position of the anchor was randomised. Each of the test OOTFs was compared against the anchor OOTF. The anchor OOTF was chosen to be that used for current SDR television production, i.e. gamma 1.2 applied on a per component basis. This OOTF is identical to the RGB_1.2.

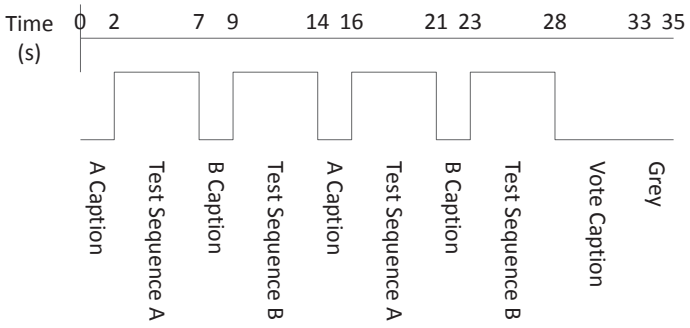

Figure 3 – Timing Diagram

A total of 19 viewers participated. The SIM2 test monitor was calibrated to have a peak brightness of 4000 cd/m^2. The monitor was then calibrated further by measuring its response to a series of standard test colours and applying a compensation process to minimise the errors from the target colours.

The files for each test were chosen from Fairchild (6), and were processed as follows:

- Source OpenEXR file is multiplied by the gain factor (iris control). The gain factor was chosen through expert viewing, on a per image basis, using the anchor image described previously (RGB_1.2),
- The OOTF under test is applied,
- The SIM2 calibration is applied to optimise the performance and colour accuracy. This processing was common for all images and formats, and

- The file is converted to a 16-bit RGB tiff file coded as LogLUV. (This conversion was made using HDRTools software suite available from JCT-VC/MPEG).

The effect of changing screen brightness, under reference lighting conditions, on the required system gamma

The test was run as a blind viewing session with 15 viewers. The tests were automated, with each viewer undertaking the tests in isolation. The test viewers completed the following test:

- The test viewer is first shown a reference image with reference peak brightness,
- The test viewer can toggle between the reference image and the test image,
- The test image has a peak brightness that differs from the reference by no more than 1 stop,
- The test viewer can adjust the system gamma of the test image, and
- Once the test viewer has adjusted the test image such that the closest perceptual match has occurred, the chosen gamma is recorded.

Choice of screen brightnesses

The instantaneous dynamic range of the human visual system has been shown to be around 13-14 stops building to greater than 16 stops if the image is viewed for more than 500 ms (Kunkel and Reinhard (11)).

The screen brightness test conditions (4000 cd/m^2 vs. 4000 cd/m^2, 4000 cd/m^2 vs. 2000 cd/m^2, 2000 cd/m^2 vs. 1000 cd/m^2 and 1000 cd/m^2 vs. 500 cd/m^2) were chosen to cover an approximate range of 18, 17, 16 and 15 stops respectively, from the darkest signal visible in the viewing environment to the peak brightness of the display.

Calculating system gammas appropriate for 4000 cd/m^2, 2000 cd/m^2 and 1000 cd/m^2 reference images

From earlier informal tests undertaken by the BBC and European Broadcasting Union (EBU) we know that an estimate of system gamma (γ – the end-to-end gamma applied between camera and monitor) can be made (Borer and Cotton(5)) based on screen brightness.

Initial expert viewing was undertaken by 10 staff members to check the validity of this test methodology and to get a better approximation of required system gamma.

During the expert viewing tests it became apparent that the suggested system gamma values given in Borer and Cotton (5) were too high, so they were reduced to those figures given in Table 2 based on expert viewings.

Screen Brightness (cd/m^2)	System Gamma
1000	1.2
2000	1.32
4000	1.45

Table 2 – Revised System Gammas for Intermediate References

Choice of images

The authors used public test sets (Frölich et. al. (7) & Fairchild (6)) to ensure that no bias was introduced in the creation of the test images. Not all features which we believe need to be tested were available in the limited number of images with which it was practical to test.

To ensure that test images covered the wide variety of high dynamic range subjects that could appear across television genres, we would need to test items such as:

- skin tone. Viewers are used to watching actors wearing stage make up under staged lighting conditions. Research suggests that skin tone without make up can be perceived as unnatural under staged lighting (Frölich et. al. (6)),
- chiaroscuro images with details in both shadows and highlights (Zakia (12)),
- specular and diffuse highlights caused by object texture (Hunter et. al. (13)) and with different spatial relationships between shadows and mid-tones,
- images with large amounts of out of focus areas which are used in drama (Stump (14)), etc.

No test images were available with areas of out of focus, all other requirements were met using these two public test image sets.

Image processing

Two sets of images are required for each combination of screen brightness and test image - the reference images and the test images. The reference images were created with an end-to-end OOTF with a reference system gamma applied to the luminance component. The test images are similarly created with a range of end-to-end OOTFs with system gammas from 0.80 to 1.60 with an increment of 0.02.

Independence of chosen output gamma to reference gamma

Since the gamma values for the reference images were only estimates, it was necessary to show that a small offset in the gamma for the reference image would not affect the change in gamma required to match images on a display with different peak brightness.

Initial expert viewings were undertaken by 10 staff members to test whether or not the chosen gamma multiplier was independent of system gamma of the reference image.

Test viewers were asked to match six image pairs in which the reference had a system gamma of 1.45 and a screen brightness of 4000 cd/m^2. The test condition showed the same image on a screen with a peak brightness of 2000 cd/m^2.

The test viewers were then shown the same six image sequences but with a different reference image system gamma (some images used 1.55, some used 1.35). The difference in the ratios between the 1.45 gamma test and the 1.35/1.55 gamma test was calculated for each image and recorded in Table 3.

If the ratio is independent of the exact gamma value used to generate the reference test images, the ratio should not change with different "reference" image gamma values.

From Table 3, it can be seen that the mean variation in system gamma ratio is low and has both positive and negative results for different images. There also appears to be no systematic offset. So it appears safe to conclude that similar gamma ratios are chosen irrespective of the reference image's system gamma.

Image Name (Reference Gamma)	Mean Variation in System Gamma Ratio
Peck Lake (1.35)	0.053
Hoover Dam (1.35)	0.034
Smoky Tunnel (1.35)	0.013
Flamingo (1.55)	-0.009
Devil's Bathtub (1.55)	0.011
507 (1.55)	-0.005

Table 3 – Calculated Difference in Results for 10 Test Candidates

RESULTS

Comparison of reference OOTF under reference viewing conditions

Figure 4 shows the difference in average Mean Opinion Scores (MOS) between the anchor OOTF and the OOTFs under evaluation; each opinion score was calculated by finding the difference between the OOTF under test and the anchor image. A positive score indicates that the viewer preferred the test OOTF over the anchor. The error bars represent a 95% confidence interval. A difference of 20 points is generally considered to be 1 grade, with positive scores indicating the test image being better than the anchor. As seen in the graph, there is a very slight preference for the Y_1.4 OOTF, however, all the error bars are very wide, and the differences are very small, therefore no defined conclusions can be drawn from the results presented in this way, which is a surprise as the test images produced with the 4 different OOTFs all looked very different.

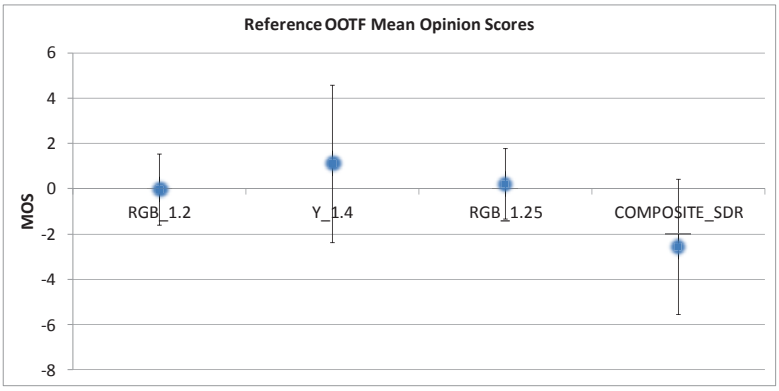

Figure 4 – Mean Opinion Scores

As noted previously, tests of this nature are prone to have a large amount of variability; As the RGB_1.2 proposal is identical to the anchor OOTF, one would have expected the error bars for that MOS to be smaller than the difference between the MOS values, but they are not. This result may indicate that the test was very hard for the viewers; therefore it may be more suitable to present the results in an alternative way.

Preference scoring

A scheme was devised in which preference scores would be allocated on a per image/per viewer basis and collated across all the tests. On a per image and per viewer basis, the difference scores for the four OOTF tests were compared for a particular image. One point was assigned to the OOTF with the highest positive difference score. In this test, the anchor is identical to the RGB_1.2 OOTF. The scheme was devised to remove any ambiguous results. An ambiguous result would arise if the reversing the polarity of the anchor vs RGB_1.2 (i.e. comparison of identical images) difference score affected the assignment of the point. The preference scores were normalised by the number of valid results per image across all the viewers. This step was required as the number of scores per image is not the same due to the removal of invalid results. The results are shown in Table 4 and plotted with 95% confidence intervals in Figure 5.

Image Name	RGB_1.2	Y_1.4	RGB_1.25	COMPOSITE SDR
AhwahneeGreatLounge	0.15	0.41	0.15	0.29
CemetryTree1	0	0.67	0.33	0
DevilsBathtub	0.11	0.33	0.3	0.26
HancockKitchenInside	0.1	0.27	0.07	0.57
M3MiddlePond	0	0.6	0.2	0.2
SmokyTunnel	0.09	0.5	0.24	0.18
st_kats9	0.06	0.69	0	0.25
stables4	0.04	0.79	0.04	0.14
TheNarrows2	0.17	0.33	0	0.5
Mean	0.08	0.51	0.15	0.27
Standard deviation	0.06	0.18	0.13	0.17
95% Confidence Interval	0.04	0.12	0.08	0.11

Table 4 – Normalised Preference Scores

The results of this experiment show that a reference OOTF Y_1.4 was judged to look the most “natural” 50% of the time, which is a significant difference when compared to the legacy SDR OOTF, RGB_1.2 (selected less than 10% of the time). Further experiments conducted after the initial experiment have shown that an OOTF with gamma 1.2 applied to luminance (not tested in these experiments) was found to perform less well than gamma 1.4 on luminance (the best performer in these tests), suggesting that ‘1.4 on luminance’ still would have been the best result even if ‘1.2 on luminance’ had also been included as a test candidate. The second experiment examines how such an OOTF can be adapted for a variety of different brightness displays.

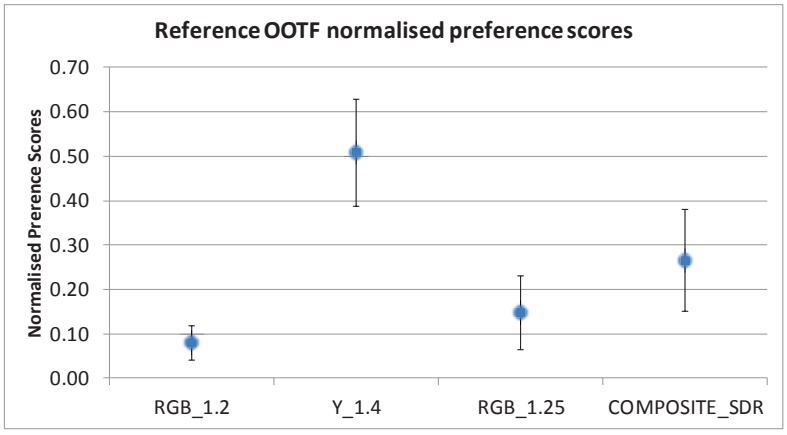

Figure 5 – Normalised Mean Preference Scores

Effect of changing screen brightness under reference lighting conditions on required system gamma

The mean chosen system gamma for all test viewers for each individual test image are shown in Figure 6 with 95% confidence intervals. The ratio between reference and non-reference system gammas, concatenated for all test brightnesses, is shown in Figure 7. Additionally, the standard deviation of 4000 cd/m^2 vs. 4000 cd/m^2 is shown. The gamma (γ) required for a screen of peak brightness L_w is given by (ITU (15)):

$$\gamma = 1.2 + 0.42log_{10}\left(\frac{L_w}{1000}\right)$$

Testing validity of results using a different screen technology

Informal tests were undertaken by a small number of BBC R&D staff using a prototype Sony BVM-X300 professional OLED monitor. Viewers were shown an image with a system gamma of 1.20 and a peak brightness of 1000 cd/m^2. They were then shown the same image with a peak brightness of 500 cd/m^2 and asked to alter the system gamma so that the images matched perceptually.

The average system gamma ratio was found to be 0.918, the equivalent system gamma ratio for the SIM2 LCD display is 1.06/1.2 = 0.88. To put these figures in context, a change in system gamma of around 0.02 is usually just perceptible to expert viewers, suggesting that the choice of display has a negligible effect on the results.

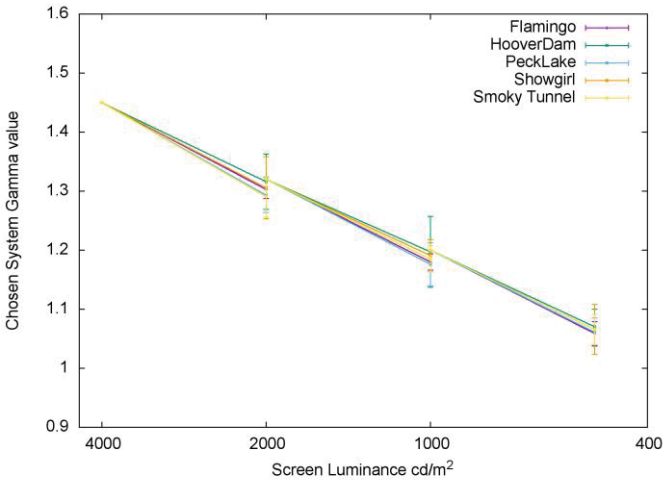

Figure 6 – Chosen System Gamma for Individual Images

CONCLUSIONS

Together, these experimental results provide the appropriate values of gamma to be used as a reference OOTF and as simple gamma adjustments in displays, which will allow the production community to create consistent high dynamic range content in the wide variety of real-world, multi-vendor production environments with different brightness displays.

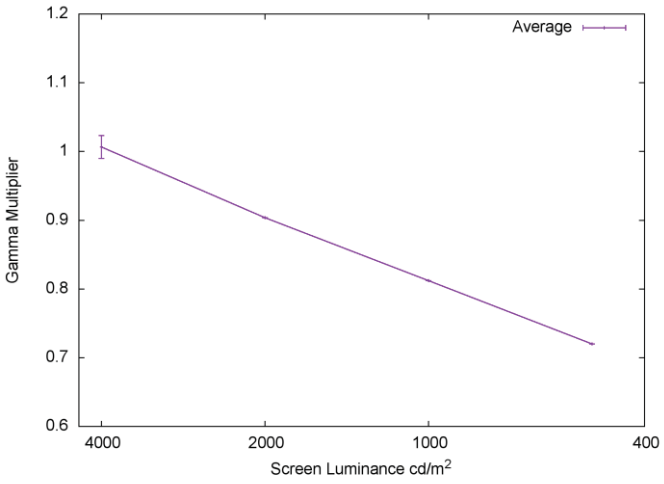

Figure 7 - Calculated Gamma Ratio for Mean of Images

The first set of subjective tests have compared reference OOTFs on the basis of realism. In absolute terms the difference in Mean Opinion Scores (MOS) for the four proponents was small (4 and 6 respectively, with a maximum of 100). It might be argued that it did not matter which OOTF was used. However, as discussed above, and supported by the reference given, it is very difficult to judge pictures without a reference. A reference image (which would be “real life”) was not possible for these tests and an anchor image was used instead. Therefore the small difference in the MOS was not entirely unexpected. However, when preferences between OOTFs were considered there was a clear preference between the candidate OOTFs. The proposal for an OOTF applying gamma 1.4 at 4000 cd/m^2 peak brightness on luminance was clearly preferred, and the result was statistically significant.

We suggest the main reason for this preference was because the OOTF was applied to the luminance part of the signal rather than to each colour component separately. Applying the OOTF to components can be mathematically demonstrated to result in changes to colour saturation and hue. By contrast, applying the OOTF to luminance does not result in such colour changes. This effect could clearly be seen in some test images and probably explains the preference for the applying gamma on luminance. Note that this effect is much more pronounced for HDR displays, which are much brighter than SDR displays and, consequently, require a higher value of system gamma.

The second set of subjective tests show that there is a logarithmic relationship between chosen system gamma (and hence gamma multiplier) and the screen brightness under reference lighting conditions. This chosen gamma multiplier is independent of the reference image's absolute value of system gamma. The equation given in the earlier BBC R&D white paper (Borer and Cotton (5)) over-estimates the required system gamma by a small margin.

More tests are required to investigate the effect of ambient lighting on required system gamma with a wider range of test images under these further "non-reference" viewing conditions. Initial tests (not reported here) showed little correlation at low ambient lighting levels with some correlation at higher levels.

Using the results of a secondary test with a different display technology, the authors suggest that the gamma adjustments are independent of the display technology, and the effects of local backlight dimming on LCD displays have only a small (if any) influence on the results.

REFERENCES

1. Hanhart, P., Korshunov, P., Ebrahimi, T., Thomas, Y. and Hoffmann, H., 2015, Subjective Quality Evaluation of High Dynamic Range Video and Display for Future TV, Journal of the Society of Motion Picture and Television Engineers, May/Jun 15, pp 1-6
2. EBU, 2014, User Requirements for Video Monitors in Production ver3, Tech3320
3. ITU-R, 2011, Reference electro-optical transfer function for flat panel displays used in HDTV studio production, BT.1886
4. Teear, I.H., 1974, Standard Conditions for Viewing in Colour TV Broadcasting, Journal of the Colour Group, vol 17, pp 43-50
5. Borer, T. and Cotton, A., 2015, A "Display Independent" High Dynamic Range Television System, BBC R&D White Paper 309
6. Fairchild, M., The HDR Photographic Survey, http://rit-mcsl.org/fairchild/HDR.html
7. Frölich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., and Brendel, H., 2014, Creating Cinematic Wide Gamut HDR-video for the Evaluation of Tone-mapping Operators
8. Ashikhmin, M., Goyal, J., 2006, A Reality Check for Tone-Mapping Operators, ACM Transactions on Applied Perception, Vol. 3, No. 4, pp 399-411
9. ITU-R RG.24, 2015, Report on the meeting of Working Party 6C, http://www.itu.int/md/R12-WP6C-C-0511
10. ITU-R, 2012, Methodology for the subjective assessment of the quality of television pictures, BT.500
11. Kunkel, T. and Reinhard, E., 2010, A Reassessment of the Simultaneous Dynamic Range of the Human Visual System, Proc. 7th Sym. on Applied Perception in Graphics and Visualisation
12. Zakia, R.D., 2013, Perception and Imaging: Photography a Way of Seeing
13. Hunter, F., Biva, S. and Fuqua, F., 2011, Light, Science and Magic: An Introduction to Photographic Lighting – 4th Edition
14. Stump, D., 2014, Digital Cinematography: Fundamentals, Tools, Techniques and Workflows
15. ITU-R, 2016, High Dynamic Range Television for Production and Programme Exchange, BT.2390

ACKNOWLEDGEMENTS

The authors would like to acknowledge Tim Borer and Andrew Cotton for their leadership in this project, Katy Noland for technical assistance and the staff of BBC R&D, NHK and the visitors to the European Broadcasting Union's Production Technology Seminar 2016 who participated in the subjective experiments.