Visually Lossless Compression of High Dynamic Range Images: A Large-Scale Evaluation

Aishwarya Sudhama[1], Matthew D. Cutone[1], Yuqian Hou[1], James Goel[3], Dale Stolitzka[5], Natan Jacobson[4], Robert S. Allison[2], Laurie M. Wilcox[1]

[1] Department of Psychology, Centre for Vision Research, York University, 4700 Keele St., Toronto, ON, Canada, M3J 1P3; +1 416-736-5659; ash10@my.yorku.ca; cutonem@yorku.ca; yhou@yorku.ca; lwilcox@yorku.ca

[2] Department of Electrical Engineering & Computer Science, Centre for Vision Research, York University, 4700 Keele St., Toronto, ON, Canada, M3J 1P3; +1 416-736-5659; allison@cse.yorku.ca

[3] Qualcomm Canada Inc., 100-105 Commerce Valley Drive West, Markham,
Ontario, Canada L3T 7W3; +1 (905) 482-5765; jgoel@qti.qualcomm.com

[4] Qualcomm., 5775 Morehouse Dr, San Diego, CA 92121, USA; +1 858-587-1121;
natanj@qti.qualcomm.com

[5] Samsung Electronics Co., Ltd., 3655 North First Street, San Jose, CA 95134, USA; +1 (408) 544 4499;
d.stolitzka@samsung.com

Abstract

High-dynamic range displays provide impressive image quality, but require markedly higher bandwidth. Here we report results of the first large-scale subjective assessment of HDR image compression. We applied the ISO/IEC 29170-2 flicker paradigm to evaluate two state-of-the art VESA codecs (DSC v1.2a, pre-release version of VDC-M) at several compression levels.

Author Keywords

High-dynamic range (HDR); Display Stream Compression (DSC); VESA Display Compression (VDC-M); visually lossless; subjective quality assessment; image compression

1. Objective and Background

The rapid progress and widespread availability of high dynamic range (HDR) and wide color gamut (WCG) technologies has increased the demand for efficient compression to support higher bandwidth across display links. Two lightweight algorithms have been developed recently to address this growing requirement. VESA Display Stream Compression standard (DSC) 1.2a is a released, lightweight video codec designed to provide low-impairment, low-latency compression for display applications with greater dynamic range [1]. The recently introduced VESA Display Compression-M Standard v1.0 (VDC-M) is a higher-complexity codec targeting higher rates of compression for use where higher pixel bandwidth is required (e.g. mobile applications) [2]. Both codecs have undergone extensive evaluation using objective error metrics such as PSNR (peak signal to noise ratio) and S-CIELAB [7]. While these methods are useful for detecting differences and failure modes, they do not necessarily predict the visibility of artefacts by human viewers [4]. Thus, in addition to objective tests, subjective trials are needed to determine if the output of a particular codec is visually lossless (observers cannot differentiate between the compressed and original content) [3]. To date, DSC1.2a has undergone extensive subjective testing [4,5] using 8-bit per color channel imagery, but not with HDR content and displays. The increased luminance dynamic range (to 1,000 nits) challenges display stream compression, particularly in high-resolution formats, and may enhance the visibility of artefacts. In this study, we conducted a large-scale subjective evaluation of DSC 1.2a and a pre-released version of VDC-M using a wide range of HDR images. The study was performed with a large number of naïve observers, at a variety of bit rates to assess the robustness of the two codecs.

2. Methods

2.1 Observers and Apparatus: All subjects (n=176) were recruited from the York University community and were screened for color vision and acuity as specified in the ISO/IEC 29170-2 subjective trial standard [3]. Ten participants were excluded based on the results of the visual pre-screening, and six participants were excluded based on their performance on catch trials (see below for details). Images were presented on one of three identical testing stations using custom scripts (MATLAB and an HDR GPU driver). Each test station consisted of a 65” Samsung TV, UN65JS9500 (3840 x2160 @ 30Hz, 60 pixels per degree at 128 cm viewing distance), an HP Z440 Workstation, (Xeon E5-1620V4 3.5GHz, 16GB RAM, 500GB SSD, Windows10-64-bit) and an AMD Radeon ™ RX460 2GB, driver Crimson ReLive edition 17.7.2 graphics card. The TVs were set to display in the DCI-P3 color space.

2.2 Stimuli: Twenty HDR images were selected from a variety of sources (Figure 1) and included content of people, natural and man-made scenery, text, and known challenging imagery [3]. The source images were non-linear 10-bit TIFF files using BT.2020 color primaries and the ST 2084 transfer characteristic curve. Eleven of the images were chosen from the Stuttgart Media University content [8] and three images were taken from each of the following sources: SJTU Media Lab [9], MIT Adobe 5K Dataset [10], and Blender Foundation [11]. It is critical to the validity of the codec evaluation that the test images present challenges to compression while at the same time avoiding impossible cases such as images similar to white noise. Therefore, as outlined below, the selection process consisted of extensive evaluation using a series of objective (PSNR, S-CIELAB [7], SSIM [12]) and subjective test procedures.

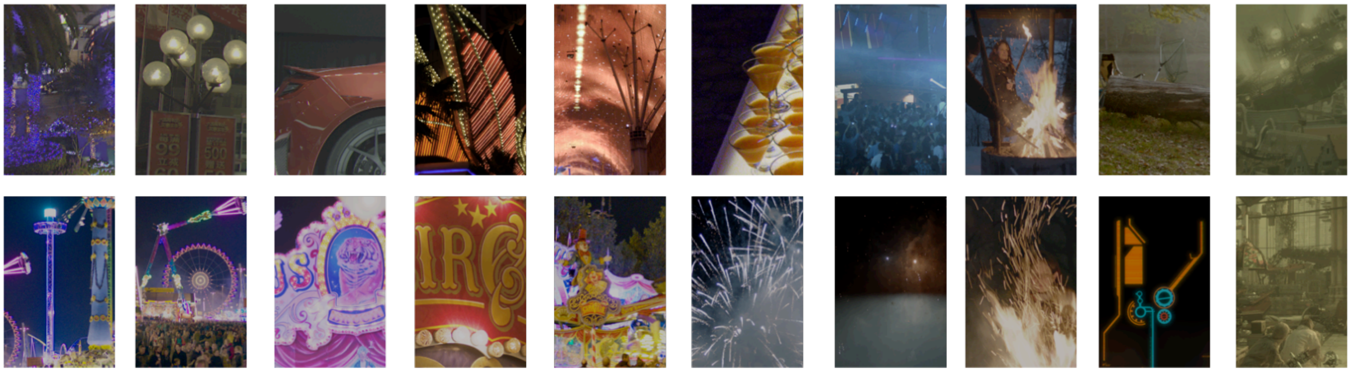


Figure 1. Thumbnails of the 20 images used to assess DSC 1.2a and VDC-M. All of the images were cropped to 400 x 600 pixels following compression and were then displayed as described in the Methods section.

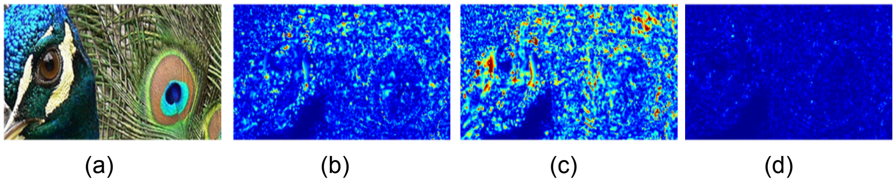
It is important to note that the ISO/IEC 29170-2 flicker paradigm applied here is not designed to assess image quality in extreme cases of lossless or lossy compression; instead, it is most informative at or near the point where artefacts are just noticeably detectable. To target these conditions, objective image analysis was used to identify image-algorithm combinations. Figure 2 shows one potential reference image (a) and its S-CIELAB heat map with color-coded error locations (b). In this case, the image-algorithm combination would be a good candidate for subsequent subjective assessment. If the S-CIELAB output resembled that shown in (c) or (d) where the artefacts were too evident or absent respectively, the image would not be further evaluated. Potential candidate images were then assessed using a modified version of the ISO/IEC flicker paradigm. Using this combination of objective and subjective assessment procedures, approximately 10,000 candidate images were culled to approximately 200 images, and then further reduced to about 50 images with image mapping and scene selection. Expert viewers used the ISO/IEC flicker paradigm to further reduce the image set to the 20 images shown above (Figure 1). VESA DSC 1.2a and a pre-released version of VDC-M compression were applied to the full frame image and a 400 x 600 pixel region was cropped from both the original (reference) and compressed version. The location of the crop was selected to constrain the observers' attention within a region of interest [5]. Each pair of images was displayed at the centre of the screen, with a background luminance of 0.48 cd/m2.

2.3 Procedure: A two-alternative forced choice (2AFC) task was performed according to the ISO/IEC 29170-2 flicker protocol (Annex B). The observer was presented with two versions of the same image (one being the compressed target and one the uncompressed reference) side-by-side, and each image alternated with the original (reference) image at a rate of 5Hz. At this rate, compression artefacts should be detectable by the human visual system in the target-original alternation [6]. The viewer was asked to indicate which of the pair contained flicker using a gamepad. Each trial was shown for 4s or until a response was made. Auditory feedback was provided (500Hz, 0.1s tone) after an incorrect response.

***2.4 Conditions**:* The two codecs were tested with RGB 4:4:4 pixel sampling. DSC 1.2a was tested at both 8 and 10 bits per pixel (bpp) while VDC-M was tested at 6 and 7.5bpp.

3. VDC-M

The VDC-M codec has an increased computational complexity relative to DSC 1.2a and is therefore able to support higher compression rates. This increase in complexity is composed of multiple factors. The fundamental coding unit size of VDC-M is an 8x2 pixel block, in comparison to the 3x1 pixel group defined in DSC 1.2a. This larger block size has allowed for several new features in VDC-M, including: 2D transform (DCT) with intra prediction, explicit block prediction, and vector entropy coding. In many cases, the 8x2 block is partitioned into smaller sub-blocks (e.g. 2x2) such that encoding/decoding steps can be performed in parallel in a hardware implementation. For SDR content, the typical use case for VDC-M is 6bpp (i.e. 4:1 compression for a 24bpp uncompressed source). In this experiment, the codec is tested at 6 and 7.5bpp for 30bpp uncompressed source content (i.e. 5:1 and 4:1).



(a) (b) (c) (d)

Figure 2. A potential reference image is illustrated in (a), a S-CIELAB lab color-coded heat map in (b) shows artefact locations. Heat maps shown in (c) and (d) illustrate cases in which the image-algorithm combination creates artefacts that are either too numerous (c) or too few (d) for use with the ISO/IEC flicker paradigm.

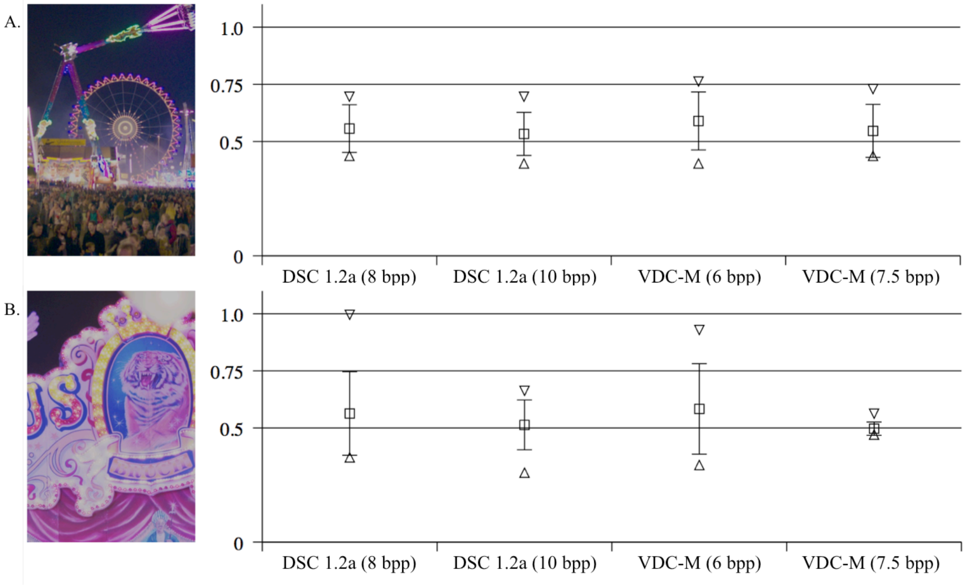
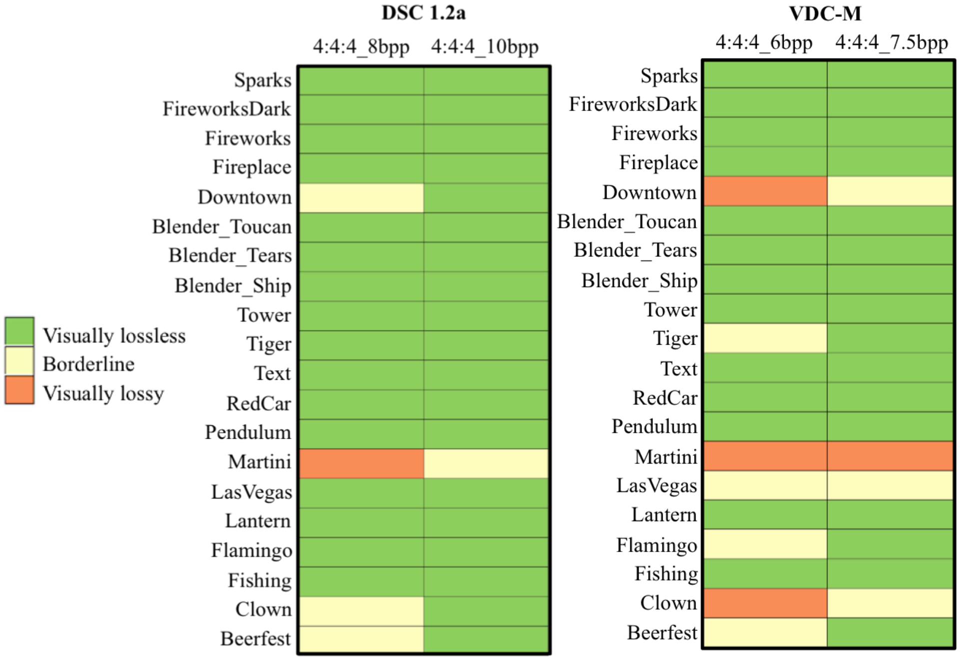


Figure 3. Sample proportion correct for two test images (A. *Pendulum,* B. *Tiger)* for DSC 1.2a and VDC-M. The test condition is indicated on the x-axis, organized by bit rate. Square symbols represent the proportion correct averaged across 10 observers. The error bars represent ± 1 standard deviation, and triangles indicate the best and worst performance. For image A. all conditions are visually lossless. For image B. VDC-M (6bpp) is borderline, while other conditions are visually lossless.



🟩 Visually lossless
🟨 Borderline
🟧 Visually lossy

	DSC 1.2a		VDC-M	
	4:4:4_8bpp	4:4:4_10bpp	4:4:4_6bpp	4:4:4_7.5bpp
Sparks	🟩	🟩	🟩	🟩
FireworksDark	🟩	🟩	🟩	🟩
Fireworks	🟩	🟩	🟩	🟩
Fireplace	🟩	🟩	🟩	🟩
Downtown	🟨	🟩	🟧	🟨
Blender_Toucan	🟩	🟩	🟩	🟩
Blender_Tears	🟩	🟩	🟩	🟩
Blender_Ship	🟩	🟩	🟩	🟩
Tower	🟩	🟩	🟩	🟩
Tiger	🟩	🟩	🟨	🟩
Text	🟩	🟩	🟩	🟩
RedCar	🟩	🟩	🟩	🟩
Pendulum	🟩	🟩	🟩	🟩
Martini	🟧	🟨	🟧	🟧
LasVegas	🟩	🟩	🟨	🟨
Lantern	🟩	🟩	🟩	🟩
Flamingo	🟩	🟩	🟨	🟩
Fishing	🟩	🟩	🟩	🟩
Clown	🟨	🟩	🟧	🟨
Beerfest	🟨	🟩	🟨	🟩

Figure 4. Heat map depicting results for conditions tested under DSC 1.2a and VDC-M compression. Orange indicates conditions where the *mean* detection rate exceeded 0.75 (visually lossy), yellow where the mean was ≤ 0.75 and the mean + one standard deviation ≥ 0.75 (borderline), and green where all detection rates were less than 0.75 (visually lossless).

4. Results and Conclusions

Results were presented graphically in the manner recommended by the ISO/IEC 29170-2 protocol. Data from all sessions and blocks of trials was collected and separated by test image and presented as proportion correct (0.5 guessing; 1.0 perfect discrimination). As illustrated by the sample data shown in Figure 3, for each condition plotted, mean proportion correct was plotted with ± 1 standard deviation along with triangles indicating the best and worst performing observers (downwards and upwards oriented triangles respectively). Figure 4 summarizes detection rates for DSC 1.2a and VDC-M for several bit rates. The ISO/IEC 29170-2 protocol recommended criteria for visually lossless performance is that no observer could correctly identify the reference image on more than 75% of the trials. As outlined previously [5] this criteria is extremely sensitive to outliers. A modified version of the criteria adopted by VESA focuses instead on the mean and standard deviation of the responses (see Figure 4 legend). For both codecs, an increase in artefact detection rates occurred systematically for the same images as bit rate decreased. DSC 1.2a was visually lossless for majority of the images tested at both bitrates. Even at a higher rate of compression, VDC-M 4:4:4 7.5bpp frequently met the visually lossless criteria.

5. Impact

We evaluated the performance of DSC 1.2a and a pre-release version of VDC-M under various levels of compression, and a large set of HDR images using the ISO/IEC flicker protocol. This is the first study to apply these state-of-the-art codecs to evaluate performance on high-dynamic range content. As described above, the HDR content was selected to challenge the codecs, in spite of this both DSC 1.2a and VDC-M performed very well. This finding is consistent with previous series of experiments using SDR images [5].

6. Acknowledgements

This work was supported by the Video Electronics Standards Association (VESA).

7. References

[1] F. G. Walls and A. S. MacInnis, “VESA Display Stream Compression for Television and Cinema Applications” IEEE Journal of Emerging and Selected Topics in Circuits and Systems **6(4)** 460-470 (2016).

[2] N. Jacobson, V. Thirumalai, R. Joshi, and J. Goel, “A new display stream compression standard under development in VESA,” Proceedings of SPIE **10396** 103960U1-103960U12 (2017).

[3] “Information technology — Advanced image coding and evaluation — Part 2: Evaluation procedure for nearly lossless coding,” International Organization of Standards, Geneva, Switzerland, ISO/IEC 29170-2:2015, (2015).

[4] D. M. Hoffman and D. Stolitzka, “A new standard method of subjective assessment of barely visible image artifacts and a new public database” Journal of the Society for Information Display **22(12)** 631–643 (2014).

[5] R. S. Allison, L. M. Wilcox, W. Wang, D. M. Hoffman, Y. Hou, J. Goel, L. Deas, D. Stolitkza, “Large Scale Subjective Evaluation of Display Stream Compression” Society for Information Display – Digest **75 (2)** 1101-1104 (2017).

[6] C. Landis, “Determinants of the Critical Flicker-Fusion Threshold” Physiological Reviews **34(2)** 259–286 (1954).

[7] X. Zhang and B. Wandell, “A spatial extension of CIELAB for digital color reproduction” Journal of the Society for Information Display **5**, 731-734 (1996).

[8] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, H. Brendel, “Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays” Proc. SPIE 9023, Digital Photography X, 90230X (2014).

[9] L. Song, Y. Liu, X. Yang, R. Xie, G. Zhai, W. Zhang, “The SJTU HDR Video Sequence Dataset”, the Eighth International Conference on Quality of Multimedia Experience (QoMEX2016), Lisboa, Portugal (2016).

[10] V. Bychkovsky, S. Paris, E. Chan, F. Durand, “Learning photographic global tonal adjustment with a database of input / output image pairs”, The Twenty-fourth IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011).

[11] The Blender Foundation. “Big buck Bunny” (CC) <https://peach.blender.org/>, “Sintel” (CC) <https://durian.blender.org/>, “Tears of Steel” (CC) <https://mango.blender.org/>.

[12] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." IEEE Transactions on Image Processing **13 (4)** 600–612 (2004).