Journal of Electronic Imaging

JElectronicImaging.org

Perspectives on the definition of visually lossless quality for mobile and large format displays

Robert S. Allison
Kjell Brunnström
Damon M. Chandler
Hannah R. Colett
Philip J. Corriveau
Scott Daly
James Goel
Juliana Y. Long
Laurie M. Wilcox
Yusizwan M. Yaacob
Shun-nan Yang
Yi Zhang

SPIE. IS&T imaging.org

Robert S. Allison, Kjell Brunnström, Damon M. Chandler, Hannah R. Colett, Philip J. Corriveau, Scott Daly, James Goel, Juliana Y. Long, Laurie M. Wilcox, Yusizwan M. Yaacob, Shun-nan Yang, Yi Zhang, "Perspectives on the definition of visually lossless quality for mobile and large format displays," *J. Electron. Imaging* **27**(5), 053035 (2018), doi: 10.1117/1.JEI.27.5.053035.

Perspectives on the definition of visually lossless quality for mobile and large format displays

Robert S. Allison,[a] Kjell Brunnström,[b,c,*] Damon M. Chandler,[d] Hannah R. Colett,[e] Philip J. Corriveau,[e] Scott Daly,[f] James Goel,[g] Juliana Y. Long,[e] Laurie M. Wilcox,[a] Yusizwan M. Yaacob,[d] Shun-nan Yang,[h] and Yi Zhang[i]
[a]York University, Centre for Vision Research, Toronto, Canada
[b]RISE AB, Acreo, Visual Media Quality, Stockholm, Sweden
[c]Mid Sweden University, Information Systems and Technology, Sundsvall, Sweden
[d]Shizuoka University, Hamamatsu, Shizuoka, Japan
[e]Intel Corp., Santa Clara, California, United States
[f]Dolby Laboratories Inc., Sunnyvale, California, United States
[g]Qualcomm Technologies, Inc., Display Video Processing Group, Markham, Canada
[h]Pacific University, Forest Grove, Oregon, United States
[i]Xi'an Jiaotong University, School of Electronic and Information Engineering, Xi'an, China

Abstract. Advances in imaging and display engineering have given rise to new and improved image and video applications that aim to maximize visual quality under given resource constraints (e.g., power, bandwidth). Because the human visual system is an imperfect sensor, the images/videos can be represented in a mathematically lossy fashion but with enough fidelity that the losses are visually imperceptible—commonly termed "visually lossless." Although a great deal of research has focused on gaining a better understanding of the limits of human vision when viewing natural images/video, a universally or even largely accepted definition of visually lossless remains elusive. Differences in testing methodologies, research objectives, and target applications have led to multiple ad-hoc definitions that are often difficult to compare to or otherwise employ in other settings. We present a compendium of technical experiments relating to both vision science and visual quality testing that together explore the research and business perspectives of visually lossless image quality, as well as review recent scientific advances. Together, the studies presented in this paper suggest that a single definition of visually lossless quality might not be appropriate; rather, a better goal would be to establish varying levels of visually lossless quality that can be quantified in terms of the testing paradigm. [DOI: 10.1117/1.JEI.27.5.053035]

Keywords: visual lossless; visual lossy; image quality; industrial perspective; mobile screen; large format displays.

Paper 170771P received Sep. 8, 2017; accepted for publication Sep. 11, 2018; published online Oct. 11, 2018.

1 Introduction

Advances in imaging and display engineering have given rise to improved and new image and video applications that aim to maximize visual quality under given resource constraints (e.g., power, bandwidth). Because the human visual system is an imperfect sensor, images/videos can be represented in a mathematically lossy fashion, but if they have enough fidelity so that the losses are not visible, they can be regarded as visually lossless. Although a great deal of research has focused on gaining a better understanding of the limits of human vision when viewing natural images/video, a largely accepted definition of visually lossless remains elusive. Differences in viewing distance and display characteristics can influence the visibilities of compression distortions. Similar arguments can be made in terms of ambient lighting; presentation time; testing paradigm; and viewers' familiarity with the images, distortions, and task.

There are a number of related terms arising from the field of compression that are now applied more broadly. For the most part, they have been used loosely, such as in discussions at conferences and standards meetings. These terms are mathematically lossless, digitally lossless, physically lossless, visually lossless, perceptually lossless, functionally lossless, and plausibly lossless.

Of the six terms mentioned, mathematically lossless refers to a comparison made directly on the code values of two possibly differing images, having been made different by one of the application processes mentioned. If there are no differences in the code values, the resulting quality is described as mathematically lossless. Digitally lossless and bit-for-bit lossless are other terms used synonymously with mathematically lossless. If the term lossless is used in isolation, it is almost always being intended as mathematically lossless. Physically lossless refers to the state of the image in the physical domain, that is, once transduced to light. As a result of display precision limitations, differences that are mathematically lossy in a digital image may still be lossless once converted to light by the display. An example is a 12 bit/color image having errors limited to the two least significant bits, and being displayed on a display with a 10-bit line driver. Other display limits include gamut limits, spatial frequency, spatial resolution limits, frame rate limit, and temporal response. Physically lossless quality can be assessed by light instrumentation equipment, and sometimes the term is used with the limitations of the measuring equipment in mind. That is, physically lossless as good as can be determined with some class of measuring instrument. The terminology of physically lossless must depend on the display or a display characterized sufficiently by a model with

*Address all correspondence to: Kjell Brunnström, E-mail: kjell.brunnstrom@ri.se

1017-9909/2018/$25.00

measurable parameters. That is, the term physically lossless quality must also include the qualifier of which display or class for which the quality is being assessed. Typically, visually lossless is next in this hierarchy of descending accuracy. It means that there may be mathematical differences, and even physical differences, but none of these are visible to the viewer once displayed. The inability to perceive these differences is due to the combined limits of the visual system and the display system. Consideration of visibility of artifacts necessitates specification of viewing conditions such as viewing distance and ambient light level. The viewing conditions are often specified for the application, such as the three-picture height viewing distance for HDTV, and an ambient consisting of <10 cd/m^2 surround and no light directly imping the display screen, such as for a Society of Motion Picture Engineering grading suite. The viewing distance is very important because there are many situations where distortions at the intended (practical) viewing distance are not visible, but can be seen if the viewer inspects the display more closely. Similar to physically moving closer to the display, the act of digital zooming the image to see distortions is not intended to be described by the term visually lossless. Windowing and leveling are tonescale operations used to view a higher dynamic range image on a lower dynamic range display, which is a terminology coming from the medical image field. The term visually lossless is also not intended to include such operations for most applications, as these are considered expert operations that the normal consumer viewer would not use, or even have available in most cases. However, such contrast boosting (via "windowing") and mean level elevation (via "leveling"), which would make an otherwise invisible low contrast distortion in a dark portion visible,[1] could be incorporated into the visually lossless criteria terminology, if sufficiently described. So, in the most technical sense, the term "visually lossless" would need to be qualified such as "visually lossless" quality at 1.5 picture heights for a UHDTV resolution standard dynamic range (SDR) (1000:1) display with a maximum luminance of 500 cd/m^2, and using a maximum contrast boosting of 2× and a mean level elevation of 20%," as an example. But rather than carry that detailed baggage, the term "visually lossless" is used more simply with the domain of the application being understood by those using the term, which is generally described in various standards documents. While that is not currently possible, when the displays' characteristics were common enough, such as in the cathode ray tube (CRT) display era, the term was practically useful. However, now with display capabilities varying so widely from the low dynamic range displays like e-paper, to the SDR displays like fixed backlight liquid crystal displays (LCDs), to the high dynamic range (HDR) displays like dual modulation, dual panel, and some organic light-emitting diodes (OLEDs), the assumption of a given display cannot be made, and must be stated or modeled. Fixed backlight is common term for a single uniform backlight that doesn't have local dimming, nor global dimming.

Perceptually lossless is often used interchangeably with visually lossless,[2] but this should be avoided since there are other perceptual dimensions, such as auditory, haptic, etc., that may be relevant to the same product, which have their own distortions for consideration.

The last two terms mentioned, functionally lossless and plausibly lossless, describe even lower levels of accuracy and are less used than the previous three, but are becoming more important with newer technologies and advanced understanding of visual quality. Functionally lossless[3] was coined to describe visible differences that do not affect the functional use of the image. Originating from computer graphics rendering, it initially referred to having lighting and geometry errors that were beyond a viewer's ability to know the difference without having the reference image available for comparison,[4] as opposed to many of the compression distortions that deviate substantially from natural image statistics and can be discerned without a reference. But it can be generalized to any case, where the visible distortions do not impact the function of the image. As an example, a histology image that contains distortions outside of the diagnostic region, such as in an empty part of a petri dish, can still be considered functionally lossless if the distortions do not impact the diagnosis. Similar examples from the aerial surveillance field are common. Last, plausibly lossless is similar to functionally lossless but tends to include changes in color, lighting, sharpness, contrast, and texture details that cannot be visually determined to be a distortion without having the reference image available.[5] A common phrase for this is that "not every blade of grass has to be in place." This concept is very important with newer techniques of image synthesis, such as generative adversarial networks,[6] but is also very important in the well-established field of color rendering. A character's clothing may be rendered a different color, but without extra information, the viewer would not be able to viably determine the distortion (loss). The plausibly lossless term is most often used when there is no task associated with the image. Both the functionally lossless and plausibly lossless criteria are often in conflict with artistic intent, and that is a complex discussion that is out of scope of this paper, so no further discussion of them will appear here.

The definition of terminology is only a first step toward understanding the issues surrounding our understanding of image quality and how best to quantify it. The goal of this paper is to provide a deeper understanding of the challenges facing the broader displays industry in their effort to provide high quality visual imagery to an increasingly sophisticated viewing public.

Human vision is important for both recognition of visual objects and guidance of visuomotor responses. Conversely, the recent emergence of diverse sizes, shapes, and aspect ratios demands both vision for recognition [e.g., ultra-high definition (HD) display resolution and HDR] and vision for action [virtual/augmented reality (VR/AR) and stereoscopic 3-D gaming]. These use cases may necessitate modification of the assessment method adopted by Video Electronics Standards Association (VESA) that emphasizes comparison of static images. VESA is an active industry trade group in the video display industry (www.vesa.org). This article reviews the applicability of such an approach and introduces alternative testing paradigms to address the multifaceted nature of display usage.

In this paper, we will give four different perspectives to the problem of defining visually lossless, illustrating the complexity of the problem, contributed by different authors, to be able to give as broad account of the topic as possible. The different sections are as follows:

- Section 2: "Business perspectives on visually lossless and lossy quality" by S. Daly.
- Section 3: "Detection of compression artifacts on laboratory, consumer, and mobile displays" by Y. Zhang, Y. Yaacob, and D. M. Chandler.
- Section 4: "Subjective assessment and the criteria for visually lossless compression" by L. M. Wilcox, R. S. Allison, and J. Goel.
- Section 5: "Usage perspectives on visually lossless and lossy quality and assessment" by H. Colett, J. Long, P. Corriveau, and S.-N. Yang.

Even with a very broad account of the problem, all issues that could affect visually lossless cannot be covered in a single article, so, for instance, aspect ratio distortions that can occur when scaling images from one production aspect ratio to fill a screen with a different aspect ratio, are not considered here. Also, any issue involving audio is outside the scope of the current article.

Together, the studies presented in this paper suggest that a single definition of visually lossless might not be appropriate; rather, a better goal would be to establish varying levels of visually lossless that can be quantified in terms of the testing application.

1.1 *Common Industrial Visual Quality Assessment*

In general, applications of visual quality occur in the industrial arena and have been directed toward a wide range of quality. This includes both testing methodologies as well as predictive models. For example, in the widely used International Telecommunication Union (ITU) guidelines for subjective video quality assessment, the Double Stimulus Continuous Quality Scale (DSCQS) method ITU-R Rec BT.500-13 (BT500)[5] or the Absolute Category Rating (ACR) ITU-T Rec P.910[7] uses a five-grade quality scale with subject input options of excellent, good, fair, poor, and bad, as shown in Table 1, to left. Another scale listed in the BT500[5] guidelines is the ITU impairment scale, which uses the following options: imperceptible, perceptual but not annoying, slightly annoying, annoying, and very annoying, see Table 1 (right column). Note that these scales were intended for a single stimulus, but can also be paired with a known reference, as in the above mentioned DSCQS, with an explicit reference or in ACR with a hidden reference. Both methods span a substantial range of visual quality, that is, they include both subthreshold and suprathreshold visible differences. In applications where lower quality images/videos are inevitable (e.g., streaming scenarios under limited or fluctuating bandwidth, or real-time compression under low-power constraints), such assessment of overall suprathreshold visual quality is exactly what is needed.

Table 1 The quality and impairment scales of BT500.[5]

Five-grade scale			
	Quality		Impairment
5	Excellent	5	Imperceptible
4	Good	4	Perceptible, but not annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying

Table 2 The comparison scale of BT500.[5]

−3	Much worse
−2	Worse
−1	Slightly worse
0	The same
+1	Slightly better
+2	Better
+3	Much better

For paired comparisons, Likert scales are often used since they have a bipolar structure that enables consideration of the two stimuli, as shown in Table 2.[5] These are generally arranged in a left-to-right orientation corresponding to two images being shown side-by-side (SBS). However, in some applications, the quality sought after is strictly visually lossless. That is, all visible differences (distortions) are designed to be below the human threshold and the intent of testing is to determine if this goal has been achieved. One can easily see that the five-grade quality scale in Table 1 (left column) has no ability to determine whether visually lossless quality occurs or not. The category "excellent" may imply visually lossless in some applications, and for some viewers, but this is generally not the case. On the other hand, the impairment scale does have the ability to assess visually lossless behavior, such as the boundary between responses 5 and 4. Likewise, the thresholds could possibly be determined from Likert scales using the responses -1, 0, $+1$, although the adjectives given are not as exact regarding threshold as does the ITU impairment scale.

In most conceptions of visually lossless, two images (or videos) are compared, with one being a reference and one being a distorted version. The distortions may not mean solely deviations from realism (artifacts, such as blocking artifacts and ringing) but include any changes from the reference, even if plausible to realism (such as color shifts, tone scale shifts, blur). Terms like original, source, and uncompressed are also used for the reference, but the reference may not always be the original version, or its source, and the distortion may not involve compression so those terms do not generalize. For example, in postproduction workflows, the term Mezzanine content is used to describe content that is compressed very lightly, but is subthreshold, and is used at certain stages of the workflow. This Mezzanine content is then further compressed for distribution. So, in this case, both the reference and distorted would be compressed video streams. Although there is not complete agreement on all of the details, the terms visually lossless, perceptually lossless, perceptually transparent, and visually identical are all referring to the same thing.

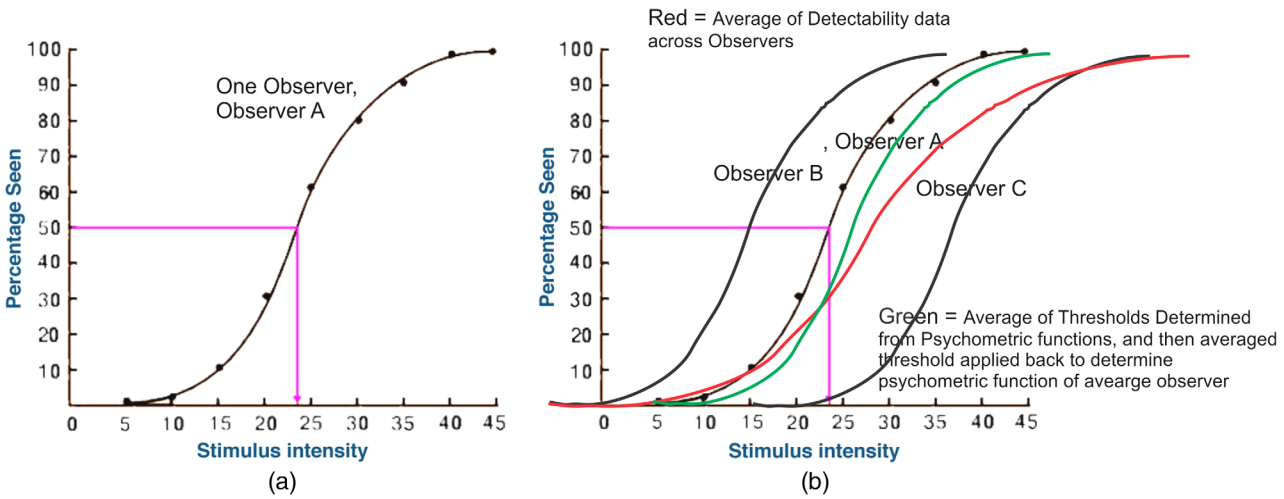


Fig. 1 (a) Psychometric function for an individual. (b) Psychometric functions for multiple subjects and different methods to determine psychometric functions or thresholds for group behavior.

1.2 *Thresholds and the Psychometric Function*

Unfortunately, the visual threshold for most dimensions of imagery is not a step function as might be implied from the impairment table in Table 1. Rather, it is a gradual transition. Rigorous psychophysical experiments (typically, vision-science experiments as opposed to visual quality testing) tend to focus more specifically on threshold perception and ignore the distinctions above threshold. A psychometric function is measured that finds the subject's probability of detection as a function of the strength of the parameter of interest, as shown in Fig. 1(a). For example, this parameter could be the contrast of the distortion or some other measure of the image's/video's physical change.

For many stimuli, psychometric functions are generally of the same shape across different individuals, but exhibit varying sensitivity [causing horizontal shifts on the x-axis, Fig. 1(b). For this example, a threshold may be assigned to the stimulus intensity corresponding to 50% seen (~24, pink arrow, left plot), but this is obviously just definitional, and then the threshold is just a shorthand for the overall position of the psychometric function. For this plot, stimuli of strengths from 40 to 45 seem to give detectability of ~100% and are just surpassing the threshold region, which may still be considered a very slight distortion. The methods used to determine such psychometric functions do not have the ability to differentiate stimuli of strengths >45, which is the suprathreshold region, to which the majority of the scales described above are allocated. To determine an average threshold across varying individuals, the detections thresholds from each are averaged and a new psychometric function can be derived, which describes the average subject [green curve in Fig. 1(b)].

One common distinction between engineering-based visual quality testing versus more traditional vision science is that the former often tests many more viewers in an attempt to gain large-scale data (e.g., for training or verification of an algorithm/design), whereas vision-science experiments typically test each viewer much more thoroughly in an attempt gain insights into human vision. In most industrial testing, there are far fewer trials per individual (sometimes just one), as well as less stimuli allocated to the threshold region, because the stimuli are needed to span a wider range of quality differences. As a result, in most visual quality testing, a psychometric function cannot be constructed per individual. But visual quality testing does have much data available as a result of testing more viewers, and attempts to determine thresholds can be made by averaging all subject responses (e.g., by looking at data for responses 4 and 5 in the ITU scale) and averaging those to get a group psychometric function.

In much visual quality testing, such as using the scales in Table 1, attempts are occasionally made to determine thresholds by averaging the responses across all observers. But without first determining the thresholds for each viewer, the overall psychometric function ends up being wider and may result in a different threshold than the average threshold determined when individual psychometric functions are measured. As a result of these many factors, experiments are generally designed to either assess the threshold or assess the full range at the expense of loss of accuracy around threshold. These design decisions involve both stimuli set as well as experimental methodologies.

1.3 *From Threshold to Just Noticeable Differences*

In most terminology, just noticeable differences (JNDs) are synonymous with the threshold corresponding to the 50% response (after correction for guessing).[8] In industrial applications, JNDs tend to be used for grouped observer perception, as opposed to describing individuals. JNDs are often added and used as a ruler to determine quality categories. For example, it has been claimed that six JNDs correspond to a difference across subjective quality categories,[9] such as from "fair to good." Another example of their usage is that one JND is not considered an advertisable difference; because it means only half the observers detect the difference. Notice that the 50% criterion is shifted from a single subject's probability of detection to the performance of a group (e.g., corresponding to the red curve in Fig. 1). Unfortunately, JND summation only works for small numbers, and saturation occurs for larger visible differences. The visual system functioning as derived from JND summation is also known to deviate from that derived from appearance estimates. For example, the luminance nonlinearity derived from thresholds deviates from one derived

from suprathreshold appearance steps (e.g., partitioning approaches). Various theories have been proposed and tested for such deviations.[10] Fortunately, for the goals of visually lossless quality, neither describing nor understanding large appearance differences is needed.

1.4 *Subthreshold Explorations*

In this century, research in quality assessment has been directed to understanding subthreshold vision. Motivations range from frustrations with the visual quality task interfering with the overall quality of experience to observations that many viewers may not be aware of visual distortions that are still considered important to the product. An example of the former is that in determining quality of experience of differing display capabilities in conveying the emotions of a narrative movie, natural viewing of the movie with audio from beginning to end is required. However, such requirements pose extreme difficulties to traditional psychophysical testing methods. The common methods of viewing, comparing, and rating video clips of 10 to 20 s duration put the viewer in a completely different state of mind than when actually watching and following the story. Examples of the latter are numerous in cases, where those involved in the professional workflow of content notice far more details relating to their craft than the consumer viewer. Rather than assuming what the viewer does not notice is not important, the presumption is that the net total of experience with the craft affects the viewer in a number of ways, e.g., honing their attention to specific attributes. These highly trained observers may be unaware of the reasons for this impact. For example, those in the craft readily use vertical camera angle placement to show dynamics of character subordination/dominance,[11] but how many consumer viewers notice such changes? Another example occurs for studies of discomfort, such as for stereoscopic displays or virtual reality (VR), where the viewer may not notice signs of impending discomfort until it is too late.

Rather than using traditional psychophysical testing (whether industrial or academic), physiological measurements can be used. They can allow for the studying naturalistic viewing, as well as the subthreshold region. Turn-key research equipment now enables eye-tracking, electroencephalographic measurements, galvanic skin responses, facial thermal emission imaging, and visible facial expression and reaction imaging. Such techniques are now currently being used to assess levels of emotional engagement as a result of technical display difference[12] or in causing stress on the oculomotor visual system, see Sec. 5 by Colett et al., below.

2 Business Perspectives on Visually Lossless and Lossy Quality

One of the key factors in favoring an accurate visually lossless descriptor as opposed to a wider ranging quality descriptor is the maturity of the technology used in the business. Businesses with mature technologies have products that are often extremely high quality, with no distortions noticeable in their product. However, they still do not want to waste effort or incur higher costs delivering a physical quality higher than visually noticeable. On the other hand, businesses with developing technologies have products, where distortions are visible, but the customer accepts that due to other factors, such as convenience, expectation level, cost, etc. In general, the developing businesses are continuously improving their technology, including cost reduction and ramping up their quality, and need to keep track of quality improvements that are nevertheless still in the visually lossy realm. As mentioned in the background, the need for visually lossless assessment or wide-ranging quality assessment will affect the distribution of stimuli, as well as the methodology, such as two alternative forced choice, paired comparisons, or comparative rating via scales. In addition to those methodology choices, the way the imagery is presented to the viewer for comparison is critical. For convenience of discussion, this section will use the term video to include digital video, digital cinema, as well as still imagery.

2.1 *Different Methods of Video Comparison*

Three key video comparison methods are sequential comparison, simultaneous comparison, and oscillation. "Simultaneous" is more generally referred to as SBS, and oscillation is more generally referred to as toggling (also as flicker). For completeness in encompassing all the methods of quality assessment, a fourth could be included, which is no comparison, that is, a single stimulus presentation (with no reference). Visually lossless in the truest sense cannot be done with single stimuli. Some distortions can indeed be assessed in a single stimulus presentation if their appearance looks entirely synthetic (e.g., blocking artifacts) or violate laws of physics (e.g., contains scene lighting incongruences due to image compositing[4]). These cases can be generalized to where the distortions' spatiotemporal statistics are inconsistent with the reference imagery statistics. However, many other distortions that are consistent with the reference imagery statistics cannot be assessed without a comparison. Examples of these include blur, contrast, color, and texture. If someone's hat changes from cyan to green as a result of a tonescale compression algorithm, the viewer would not be able to detect that difference without a comparison image, since both colors are plausible to a third-party viewer. A better term than visually lossless for the indistinguishable distortions as assessed by single stimulus testing is plausibly lossless.

For the traditional test video clips of 10 to 15 s duration, it is known that it is much easier to see differences when the video clips are shown SBS than when they are shown sequentially. A recent study verified this by directly comparing the two methods.[13] The experiment was identical for both cases, including display, stimuli, and task. The experiment tested one parameter of display capability: maximum luminance for HDR. In the sequential testing, one Dolby professional reference display (pulsar) was used. For the SBS testing, two pulsar displays were used. The resolution of each was full HD (1920×1080), the diagonal was 42 in., the bit-depth was 12 bits red, green, blue (RGB), the color gamut of the signal was 709, the black level remained constant at 0.005, and the ambient was 20 lux. A hidden upper anchor was used for each comparison. The viewer's task was to rate the quality (according to their own personal preference) of each of the two stimuli shown using a Likert scale. The maximum luminances tested were 100, 400, 1000, and 4000 cd/m^2. Six different HDR video clips were used, where two different max luminances were compared in each trial. The main conclusion of the results (shown in Fig. 2) is that sequential comparisons are more difficult than the SBS.

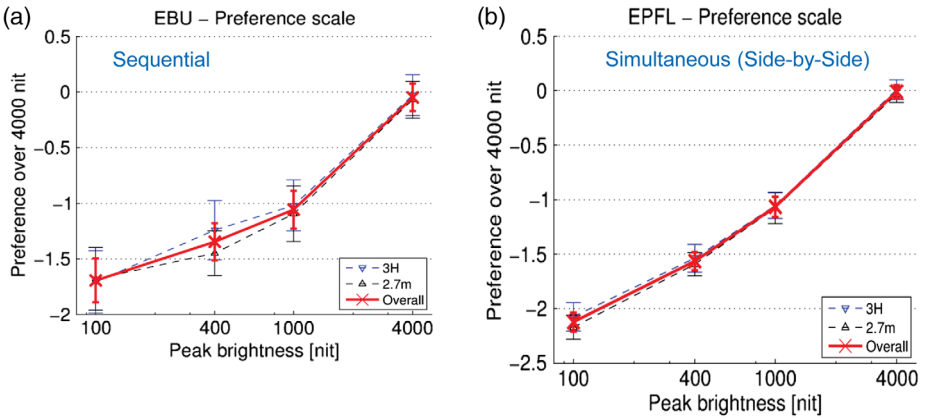


Fig. 2 Comparison between (a) sequential versus (b) side-by-side comparisons for the same stimuli, displays, and subjective task (preference). Although this is not a quality scale, there exist well established techniques to convert pair comparison preference data into quality score, such as MOS.[14] The sequential testing results were conducted by the EBU while the SBS were conducted by EPFL.

This shows up both in terms of the confidence intervals and the shape of the curves. The confidence intervals are clearly seen to be on average 2× larger for the sequential comparison task, and the range of quality is reduced. For example, there is not a significant distinction between the 400 and 1000 cd/m^2 versions in the sequential testing, while there is a clear distinction across all four tested stimuli parameters in the SBS methodology.

To better understand why the SBS comparisons give more pronounced quality distinctions, it is worth noting that any image comparison requiring a viewer's response is a task involving various stages of visual memory and mental mapping. Figure 3 shows some of the key processes for the rating comparisons as used in the mentioned experiment. Both of the compared stimuli cannot be foveated at the same time and, thus, a reason to use the term SBS over the term

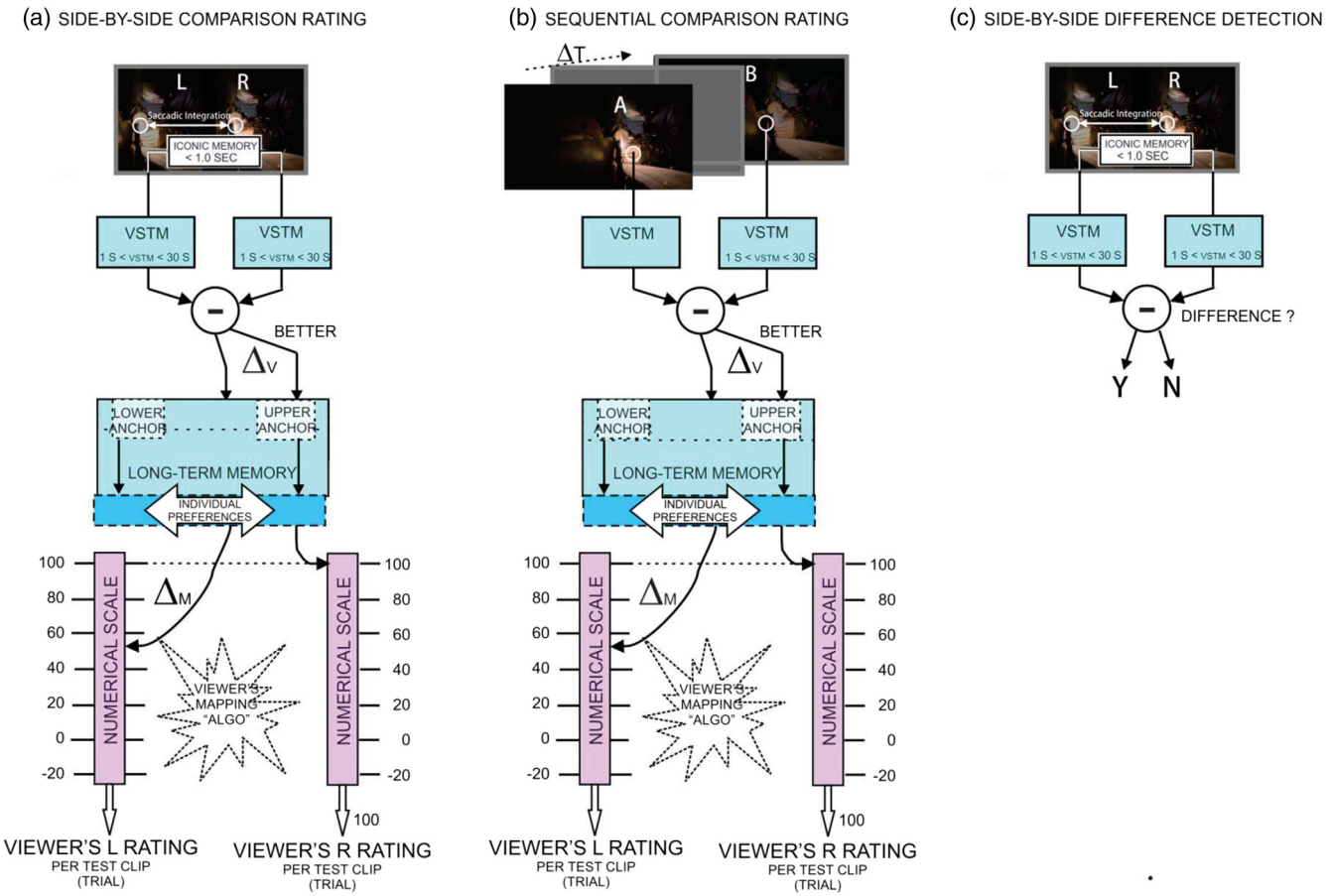


Fig. 3 Key memory and mapping stages for (a) an SBS rating task, (b) a sequential rating task, and (c) an SBS visually lossy detection task. Note: sequential is referring to viewing one entire test video clip, followed by another one (the other half of a pair with differing parameters).

simultaneous, so in the SBS method (leftmost plot), saccadic eye movements are required to compare the left and the right stimuli. Iconic memory is the term for the portion of visual memory that integrates imagery across saccades and enables us to build up a mental picture of the world having a wider field-of-view (FoV) than the fovea's mere 4 to 6 deg.[15,16] In the SBS methodology, the iconic memory is used for an additional purpose than building up a mental image; it is also used for comparing similar image regions. Regardless of its end purpose, it is still limited to be less than 1 s. These visible differences are registered in the visual short-term memory (VSTM), and its duration limits come into play.[17,18] These can be considered to hold the visible representations in the range from about 1 to 30 s. This upper limit suggests why video clips of duration less than 15 s are preferred in the testing community. The visible differences, ΔV, are noted from those in the VSTM. To go from these visible differences to a subject's numerical rating, these visible differences must be mapped into that rating range.

This requires memory of previous stimuli being shown, which would have occurred further back in time than the limits of the VSTM. In addition, if upper or lower anchors are not used (the experiment in Fig. 3 had only an upper anchor), long-term memory of video quality over perhaps years or decades may be involved. Further, individual preferences on which image features are more important (contrast versus color versus sharpness versus texture, etc.) act as biases on the long-term memory. Last, from this internal range of magnitude of visible differences, visual quality must be mapped into a numerical scale. This involves higher level cognition than the previous steps and is susceptible to even greater subject variability. To no surprise, the higher accuracy memory functions have the shorter durations. So, in terms of accuracy, the iconic is best, followed by VSTM, and then long-term excluding rare eidetic individuals. The case for sequential comparison is shown in the middle. The temporal delta would be greater than 10 to 15 s for typical video quality testing. That methodology deprives the visual system of the iconic memory being able to input localized visible comparisons to the VSTM, because many foveations to different portions of the image would have occurred before the other paired stimuli is seen. That is the most likely source of the larger confidence intervals and range compression in Fig. 3 for the sequential method.

Let us now consider the binary task of assessing visual fidelity (i.e., whether something is visually lossless or lossy), as shown in Fig. 3(c) for a SBS comparison. Since there is no rating required, a simple yes or no response can be given. Thus, the task removes the inaccuracy and biases of long-term memory, as well as individual variations in mapping their visual memory to a rating scale. Fortunately, for the businesses, where visually lossless is the most relevant criterion, their use of experiments designed around a visually lossless criterion is able to obtain much more consistent and accurate data.

The third approach mentioned, toggling, reduces the internal processing and memory load of the viewer even further. Toggling has been used since digital imaging systems with frame buffers were available in the late 70s. The term comes from a toggle switch, and the technique is still commonly used by image processing algorithm developers to look for differences in their resulting images. It is generally used for still images. It has also been used for video clips, but with less success. The two images to be compared are displayed in register (i.e., to the exact pixel position) on the screen, and the viewer toggles as desired between the two images . In current systems, left or right keyboard arrows are often used to swap (or toggle) images being displayed, as well as the space bar. The toggle switch traditionally had two positions and allowed instantaneous swapping of its inputs, and these features are preserved with the newer methods, such as using a keyboard. Occasionally, toggling is referred to as sequential when still images are toggled, but the majority of work in this field does not use sequential to refer to the rapid alternation used in oscillation or toggling. The change occurs in-place and with no interstimulus interval or blanking field, which might cause masking. Spatial and amplitude differences thus pick up an additional temporal modulation. Differences that would previously be below threshold using just SBS comparisons often become visible. This occurs for several reasons. One is that detecting visible differences in an image requires a search over the two compared images for differences. It can take a substantial amount of time to scan and foveate an entire image, particularly for detailed imagery that may be displayed with a FoV as large as 67 deg (4 k display viewed at the specified distance of 1.5 picture heights). The imposed temporal modulations caused by toggling enables better detection in the periphery (which while having poorer spatial resolution, has better temporal bandwidth and sensitivity than the fovea), aiding the viewer to find and then foveate regions formerly in the near or far periphery. So, the toggling substantially aids the search task. In addition, the lack of needed eye movements for SBS comparisons (once a region having difference is found) aids in the detection of small spatial phase shifts that would be lost across a saccade. A third reason is that even in the fovea, the addition of temporal modulation at the right frequency can improve detection. Figure 4 shows the spatiotemporal contrast sensitivity function (CSF). The temporal frequencies caused by toggling can shift the spatial frequencies of the distortion to a more sensitive part of the CSF as compared to what occurs with a static image comparison (shown in general on the left). While the highest spatial frequencies do not change that much between the two cases, there is a noticeable change at the frequencies near the peak, and a substantial change for spatial frequencies that are lower.

While toggling was originally an ad hoc technique, it has recently been made more rigorous[19] by removing the viewer's control and having the images automatically oscillate in place at a specific frequency. For the CSF at the light adaptation level shown in Fig. 4, it can be seen how an oscillation of 5 Hz maximizes the sensitivity to all visible spatial frequencies, as compared to a static, or still image comparison. Since the eye does not hold steady when foveating a region (there are always drift eye movements), the temporal frequencies for a static image comparison are not at 0 Hz (Hertz). An estimate of the temporal frequencies involved for static image viewing is shown as around 0.11 Hz in the diagram, although it is better to describe these drift eye movements in terms of velocity. The difference between the 5 and the 7.5 Hz, as suggested in Ref. 19, is relatively minor and a change in CSF light adaptation level going upward in cd/m^2 would likely put the 7.5 Hz value on

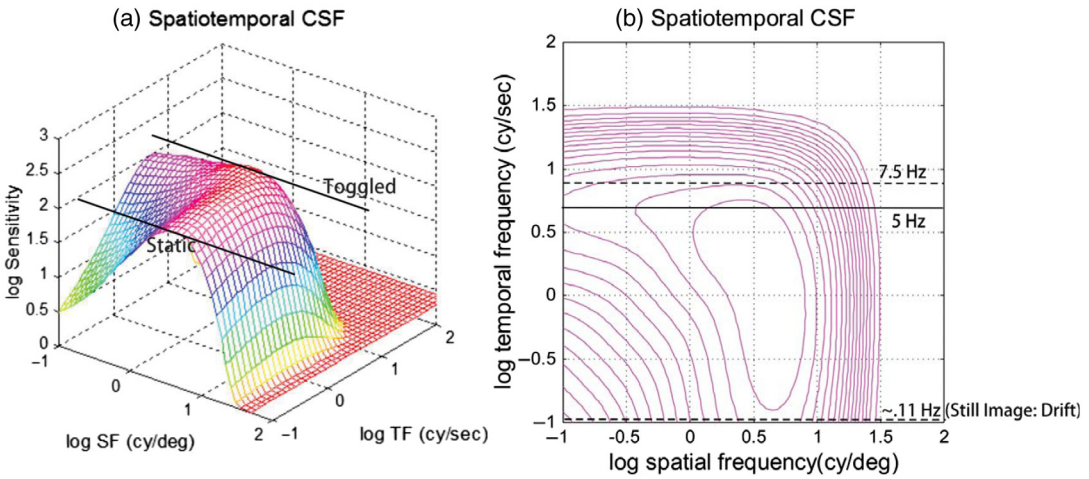


Fig. 4 Spatiotemporal CSF (at ~light adaption level of 10 cd/m2) showing (a) the general effect in a surface plot and (b) more specific changes in sensitivity in the contour plot for the oscillation techniques. Contours deltas are 0.25 log10 in sensitivity task (preference).

the CSF peak and ridge. A related approach for imposing motion on distortions to make them more salient has been used for studying amplitude quantization by phase shifting the quantization interval as a function of time.[20] These techniques result in the best ability of the visual system to see differences, and can also speed up the search time, but may not be relevant to the business application as will be described later.

2.2 *Calibration to the Display*

Calibration is needed because, while it is possible to determine the contrast required for detection of a given frequency component of a distortion, the contrast per code value depends on the luminance calibration [generally referred to as the display's electro-optical transfer function (EOTF)]. Increasing a display's contrast and using the same signal quantization results in an increase in the contrast per code value. If that increase is large enough, a previously subthreshold frequency will become visible. A recent example of this is that the increased dynamic range of HDR displays required an increase from the previously acceptable 8 bits/color to 10 bits for consumer usage and 12 bits for professionals. Similar phenomena also occur for the other image and perceptual dimensions listed above. Of the various visual behavior relating to thresholds, masking is the most impervious to lack of calibration, since once it rises above absolute threshold (i.e., no masking), it almost follows a linear signal-to-noise ratio behavior. For systems where color, dynamic range, resolution, frame rate, etc., are approximately fixed, then prediction of masking can provide a strong visual foundation for quality prediction, such as shown by uncalibrated models.[21–23] However, most display ecosystems are moving away from that situation and are trending toward more variability along these key display capability dimensions. At present, current visual models that can be calibrated to calibrated displays[24–26] have been shown to perform better in cases where display capabilities are not fixed, such as HDR.[13]

While there were many businesses unable to design for visually lossless quality, there were niche applications, where it was indeed possible to quantify most of these parameters, or at least limit them to specific ranges. This particularly occurred in closed systems, where the product included the display, the proprietary image format, and the encoding. Examples of these include some defense imaging systems (e.g., aerial image analysis), some medical systems, high-end graphic arts WYSIWYG (what you see is what you get) systems, and cinematic postproduction. In other applications, while there were some unknown calibration dimensions, visually lossless criteria could be used in the design by assuming standardized specs and ideal or worst-case parameters (such as three-picture-height viewing distance and a specified EOTF for HDTV[27]). For handling the unknowns of display reflectivity and ambient light, which have a strong interaction on the black level, techniques like the picture line-up generation equipment (PLUGE) were developed. PLUGE signal is a greyscale test pattern that can be used to adjust the black level and contrast of a picture monitor.

Fortunately, the current trends are that the display is becoming more knowable and quantifiable, and thus enabling closer adherence to visually lossless goals. For one, the displays are much more stable than they have been in the past, especially TVs (televisions), which had much thermal drift causing color and convergence errors in the CRT era. More importantly, there are standardized pathways for the display to communicate its capability to the delivery system. As an example, extended display identification data metadata that are exchanged from a display to a graphics card [and advanced services that deliver media over the Internet directly to the consumers without using a broadcast, a cable or a IPTV network, so-called advanced over-the-top (OTT) services] contain information about the display's primaries, its tonescale EOTF,[28] of which gamma is a legacy example,[29] and its pixel resolution. More advanced metadata are now being used in a number of applications, where these values are augmented by the minimum and maximum luminances, bit-depth, and other parameters of the content.[30] Further, dynamic metadata are being used to pass essential signal information to the display in order to aid tone-mapping and gamut mapping algorithms, motivated because the color volume of displays can now vary so substantially.[31] Ambient light sensors are becoming more advanced, having V_λ sensitivity to match the eye, and can be used for display's internal algorithms to tailor

the signal to the resulting black level changes. Even the key weakness in spatial calibration, i.e., the viewing distance, has a pathway to be solved with presence detectors (motivated by energy conservation) and depth sensors (motivated by interactivity), which are making continual headway into display products. Finally, the burgeoning head-mounted displays for VR have the fortunate advantages that the viewing distances are exactly known (as designed for in the optics) and the ambience can be easily controlled (generally kept dark). Thus, the video content delivery system can tailor the signal sent to the display so that the advanced visual model approaches aiming for visually lossless quality, which require such calibration, can finally be used to their theoretical intention.

2.3 *Business Considerations*

A famous sign in many service businesses is "cheap, fast, good—pick any two." It is likely obvious to any reader that increasing quality comes with a cost. In display hardware, there is a general struggle against physics to increase quality, offered initially at a higher cost, and then gradually the manufacturing efficiencies and scale of production can bring the costs down. Similar constraints are involved in the compression and video chip business. Rarely does one see a quality improvement and a cost reduction being introduced at the same time. For those wanting both, they must wait and essentially be late adopters. In this section, we will start with an anecdotal example so that concrete details can be discussed, and then, we will describe some general issues.

The plot in Fig. 2 was from an experiment[13] to provide data on whether the TV industry should develop a new ecosystem for HDR. There are a number of key attributes involved in HDR, including bit-depth, black level, local contrast, mid-tone contrast, compression technique, average luminance level, and maximum luminance. While HDR includes increasing the range at the dark end as well as the bright end, one of the unique attributes of HDR is more accurate rendering of highlights than traditional video. Such highlights include both specular reflections as well as emissive objects (visible light sources) and can require very high maximum luminance.[32] A study was designed to specifically probe this aspect in comparison to existing TV standards, known as SDR, and standardized in ITU-R Rec BT.709,[33] with an EOTF subsequently defined in ITU-R Rec BT.1886.[27] Most viewers watching SDR see only 8 bits/color video that is compressed. One aspect of HDR is that it requires a higher bit-depth than SDR, and details of whether 10 or 12 bits/color are needed depend on viewing conditions. Currently, in television systems, HDR is generally bundled with an increase in spatial resolution and color gamut as well, for example, to going from the BT.709 (sRGB) color gamut to the DCI P3 (Digital Cinema Initiative) gamut or even wider with the ITU-R Rec. BT.2020 gamut.[34] But in order to focus solely on the parameter of maximum luminance, the study used uncompressed videos at 12 bits, all with a BT.709 color gamut and an HDTV pixel dimensions (1920×1080). Four maximum luminance values were studied. They were placed approximately on a logarithmic luminance scale based on general visual system properties. The four luminances were 100, 400, 1000, and 4000 cd/m^2. Deviations from strict logarithmic spacing were motivated by practical existing television systems and displays.

The existing SDR TV system was designed for $\sim$100 cd/m^2 as the maximum. In many systems, reference white, which is generally the diffuse white maximum luminance is set to 100 cd/m^2 and the peak luminance (the maximum luminances) is set to 120 cd/m^2 and in calibrated studios, the reference monitors are set very close to this value. This is true for both episodic and live broadcast video content, and is the maximum luminance that is seen by individuals involved in the approval process (cinematographer, colorist, director, producer for episodic content, and the video shader and producer for live content) before distribution occurs. The ambient lighting followed the industry production specs of producing a surround of 5 to 10 cd/m^2. The next value, 400 cd/m^2, was selected as a typical higher-end consumer TV max luminance at the time of the study. As a reminder, the content seen at 100 cd/m^2 by the approvers is generally stretched upward in most TVs. The value of 1000 cd/m^2 was selected to represent the capability of the first generation of consumer HDR TVs. Last, the 4000 cd/m^2 value was selected because that was the maximum luminance capability of the professional HDR displays used in the experiment.

Initial attempts at using the BT500 five-point rating scale (excellent, good, fair, poor, bad) in pilot studies were inconclusive because a majority of viewers rated the lowest capability value (100 cd/m^2) as excellent, and there was no headroom on the scale to indicate higher quality than that. This was partially due to their inexperience seeing uncompressed 12-bit video, as well as a reference display (such as having lower noise, better uniformity, etc.). As a result of lack of useful guidance from the BT500 document, it was decided the experiment needed to explore testing options as well as the maximum luminance parameter. Two key comparison methodologies were agreed upon, a sequential and a SBS comparison. Video clips of 10 to 15 s were used based on common video testing, so the sequential method meant that one version of a video clip was shown, followed by a version with a different max luminance, all being shown on a single HDR reference display, and then followed by the viewer's rating. For the SBS testing, two identical calibrated displays were arranged so that viewers could compare both at the same time and arranged so each was seen with an approximately orthogonal viewing angle to the display screen. This approach has traditionally been avoided for rigorous studies in the past due to difficulties in getting two displays to have the same color, tone scale, and black level. However, modern digitally driven reference displays with internal light sensors, thermal regulation, and compensatory image processing can enable such displays to appear identical. Randomization of various contents with known parameters was used, in case there might be a small physical bias, despite being physically immeasurable. After presentation of the video test pair, the viewer was asked for a preference rating comparison. For the SBS testing, the relative quality rating scale shown below was used. For the sequential testing, it was modified to replace L and R with A and B, where A was explained to be the first instance of the sequentially shown pair, see Fig. 5.

The results have been discussed earlier in this section with the SBS having better confidence intervals than the

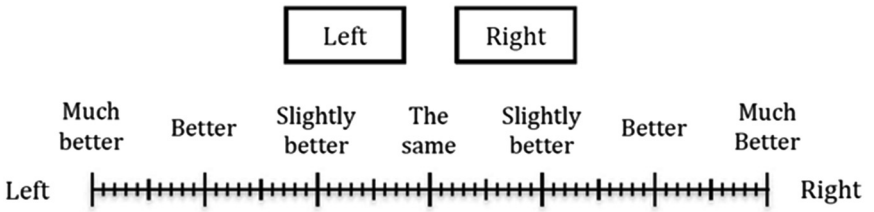


Fig. 5 An example of the Likert scale used for SBS testing, in place of the BT500 5-point rating scale, as described in the text.

sequential, as well as having a larger range of preference. What visual memory and cognitive processes are involved in each methodology have also been discussed. Let us now discuss some key business aspects. For a new television ecosystem, both the televisions and the video signals need to be updated. These involve two key different industries: the television set manufacturers and the broadcasters. For television makers' customers, the majority of TV sales involve SBS viewing of competing TV products arranged in a store at the time of the purchasing decision. Some customers may be influenced by written ratings, descriptions, and recommendations in either mainstream or more technical press, but most of the time, a SBS viewing is involved. The broadcasters have a different situation since it is generally not possible for their consumers to view their service compared to a competitor's (e.g., a different network) in a SBS manner. Rather, comparison is made by the consumer in a sequential manner by changing the channel.

The plot in Fig. 2 shows that the viewers using sequential comparisons were not able to show preference differences for the 400 and 1000 cd/m^2 parameters confidently. This is very important for business considerations in 2015 to 2017, as HDR TVs are being introduced. SDR TVs are typically 300 to 500 cd/m^2, and the first generation of HDR TVs is typically around 1000 cd/m^2. The sequential testing does not give any confidence to the preference of the new 1000 cd/m^2 HDR TVs over the current SDR TVs, while the SBS testing does give substantial confidence. The sequential results directly relevant to the broadcasting business would not be able to indicate with confidence that a change to a 1000 cd/m^2 system would be worthwhile, whereas the SBS results that are directly relevant to the TV set makers does conclude with confidence that change would be preferred to the viewer. However, both businesses involved in the ecosystem need the other business to agree to a similar upgrade. Assuming the trend of increasing maximum luminance can continue and ranges closer to 4000 cd/m^2 will eventually be reached, a future-oriented decision might be for both business segments to agree to move forward with HDR. Another way to look at the results, however, is that the SBS gets closer to the true perceptual experience of the viewer, whether or not they can see the comparisons directly in actual application. Of course, a critical customer of many broadcasters is the advertising industry, and their professional viewers would likely be able to see SBS comparisons in a production suite. As a result of these many factors, the broadcasting industry in several key regions decided to go ahead with HDR transmission. It is not clear if it was the future capability considerations or the benevolence to the viewer that was the dominating factor.

General business considerations regarding visually lossless or lossy quality are specific to the business. For example, visually lossless criteria are relevant for mature businesses already delivering a high quality, with examples being those that have a six-sigma defect strategy.[35] Visually lossless is also relevant for businesses with high-end products and high cost ranges. Examples in printing and video include most of the production workflow. An example of visually lossless compression includes what is known as mezzanine compression, having low compression ratios, below 2-3:1, and yet still use advanced techniques like wavelet or discrete cosine transform based compression. Businesses, where visually lossy quality ratings are more relevant, include newly developing businesses, developing products offering new features and conveniences, and businesses specializing in lower-cost products. For example, new businesses arising to compete with mature businesses usually begin with a lower quality and increase it as they expand their market. Streaming is a good example of a service business that initially had very low quality (circa 2006), whose quality weaknesses included not only the customers' bandwidth but also color and tone miscalibration. Now, however, there are streaming services of the highest quality, with 4k resolution at 10 bits/color and visually lossless performance for three-picture-height viewing distance.

For the businesses where visually lossless is the most relevant, each of the three comparison methods is suited toward different applications. Toggling (in particular, the automatic alternation techniques known as flicker) is most suited to imaging applications that are information-task based, where small features and minute phase shifts may be important, and the localization shortcut aspect of toggling can be a surrogate for a strenuous search process, in particular when it is unknown which elements of the imagery are most critical to the task. Examples include products and services for forensic, histology, aerial imaging, scientific visualization, medical, etc. A special case is for products within the video path, where the customer is a technical person that uses such a toggling technique for assessment, even if the end customer of the entire video path is a nonexpert consumer. Applications where results from SBS testing are most relevant include products that are generally purchased in stores, and competitor products are available. Televisions fall in this category as well as mobile displays to a lesser degree. Last, applications where sequential testing methodology is most suited include most consumer services, such as broadcast, cable, and internet delivery (i.e., OTT) of video. However, particular companies aiming for the highest levels of quality may decide on one of the other methods if their philosophy is to deliver the best quality to their customer (even if the typical customer does not notice it; see

physiological testing discussion in the background for such motivations).

3 Detection of Compression Artifacts on Laboratory, Consumer, and Mobile Displays

As outlined in Secs. 1 and 2, a range of parameters have been evaluated in threshold-based approaches to quality assessment (using forced-choice procedures and calibrated displays). However, practitioners often find that such thresholds are much lower than commonly visible in many applications, particularly when display characterization is not performed (see Wilcox et al. later in this paper). In addition to the impact of the task demands, three candidates for such discrepancy are: (1) the display, (2) the signal, and (3) the viewing distance/angle. In the case of the display, factors such as contrast loss due to tonescale variations, ambient light, and display reflectivity, motion blur due to temporal response, loss of high frequencies due to spatial modulation transfer function (MTF) and dynamic range variations are considered the most likely to influence thresholds. Regarding the signal, the content noise level and texture are the primary suspects in elevating thresholds due to masking. Last, because psychophysical thresholds have a strong frequency dependence, viewing distance underestimation can shift expected frequencies to higher values, where the thresholds are generally higher. Off-angle viewing can also significantly lower the contrast displayed with LCD technologies, thus lowering the contrast of the distortion from that expected using optimal threshold data.

Consequently, it remains unclear whether such thresholds are valid when measured for true broadband compression distortions in actual images/videos presented on mobile and consumer-grade displays. In this section, we discuss our explorations of the display portion of the issue. Specifically, we asked the following:

1. Can thresholds measured on mobile devices yield the same results as those measured on laboratory and desktop displays when viewing conditions and display EOTFs are kept constant?
2. How are the thresholds affected when EOTFs change on mobile displays, and do such changes agree with model predictions?
3. How do the variabilities in thresholds due to (1) and (2) compare to the variability across subjects, content, and gaze location?

Here, we present some preliminary findings of a pilot experiment designed to shed light on these issues. We measured contrast detection thresholds for high efficiency video coding (HEVC)[36] distortions in small images using a mobile device (Apple iPad), and a forced-choice procedure. We discuss how these thresholds compare to similar thresholds measured on other displays, on the same display but with different display settings.

3.1 *Quantifying HEVC Distortion Visibility via Contrast Detection Thresholds*

As we mentioned in Sec. 1.2, one candidate definition of visual losslessness is the inability of a human subject to visually detect the changes (distortions) resulting from compression. If the compression distortions are indeed below the threshold of visual detection, then the viewer would not be able to distinguish the distorted image/video from the original image/video. In terms of image quality, the distorted image would be of equivalent visual quality to the original. (Indeed, it is possible to achieve equivalent quality even if the distortions are visible; see, e.g., Ref. 37. Nonetheless, detection thresholds can represent conservative estimates of quality equivalence.)

An important question when defining such detection thresholds is how to quantify the physical magnitude of the distortion. Early threshold measurements were made in terms of quantization step sizes or peak-signal-to-noise ratio (PSNR);[38] however, these are digital rather than physical measurements, and particularly for quantization step sizes, the resulting physical distortion for a given image can change significantly depending on the image and display.

To overcome this limitation, other researchers have quantified the distortion in terms of its physical contrast, following from classical contrast detection studies from the visual psychophysics literature (see Ref. 39 for a review). In such classical studies, there is a target of detection, and there is possibly a masking pattern (commonly referred to simply as a mask) upon which that target is presented. Numerous studies have measured contrast thresholds for visual detection of targets consisting of sine-wave gratings, Gabor patches, bandlimited noise, or other simple patterns. These experiments have been conducted both in the unmasked paradigm in which the target is placed against a blank background; and in the masked paradigm using masks consisting of sine-wave gratings, Gabor patterns, noise, and some natural images.

For compressed images, the compression distortions are considered to be the target of detection, and the undistorted image is considered to be the mask upon which the distortions are placed. Figure 6 illustrates this mask + target paradigm. The compressed image, which is shown in Fig. 6(a), consists of two components: (1) the compression distortions which serve as the target of detection, as shown in Fig. 6(b) and (2) the original (uncompressed) image which serves as the mask upon which the distortions are presented, as shown in Fig. 6(c).

Previous studies employing distortion-type targets have used root mean square (RMS) contrast as the contrast metric, which is defined as follows for (mean-offset) target $\mathbf{t}$ presented against mask $\mathbf{m}$:

$$C(\mathbf{t}|\mathbf{m}) = \frac{1}{\mu_{L(\boldsymbol{m})}}\left(\frac{1}{N}\sum_{i=1}^{N}[L(t_i)-\mu_{L(\boldsymbol{t})}]^2\right)^{\frac{1}{2}} = \frac{\sigma_{L(\boldsymbol{t})}}{\mu_{L(\boldsymbol{m})}},$$

where $\mu_{L(\boldsymbol{t})}$ and $\mu_{L(\boldsymbol{m})}$ are the mean luminances of the target and mask, respectively; where $L(t_i)$ is the luminance of the i'th pixel of the target and where N is the total number of pixels in the target. The RMS contrast is the standard deviation of the target's luminances normalized by the mean luminance of the mask. Note that when measuring the RMS contrast of the distortions within a distorted image ($\mathbf{d}$), the target $\mathbf{t}$ is computed from the distorted and original images via $\mathbf{t} = \mathbf{d} - \mathbf{m} + \mu_{\boldsymbol{m}}$, where $\mu_{\boldsymbol{m}}$ is the mean pixel value of the original image, followed by clipping to the 8-bit pixel-value range, if necessary. Thus, as shown in

Fig. 6 An image compressed with JPEG compression (b) can be envisaged as an original image (a) to which distortions (c) have been added. The goal is to detect the target (distortions) in the presence of the mask (image).

Fig. 6(c), the target is actually a mean-offset version of the distortions.

3.2 *Effect of Display Type: Mobile vs. Desktop vs. Laboratory*

Contrast detection thresholds for HEVC[36] compression distortions were measured for crops from two images from the CSIQ masking database;[39] images Shroom and SunsetColor (see Fig. 7). The compressed images were generated by using the reference HEVC encoder and by adjusting the quantization parameter value from 1 to 51.

The thresholds were measured on three displays:

- a display++ LCD monitor from Cambridge Research Systems,
- a consumer-grade LCD monitor from I-O Data, and
- an Apple iPad Air 2 (a tablet small enough to be considered a mobile device).

All three displays were adjusted to have similar EOTFs. The EOTFs were measured by using a DataColor Spyder5 in a darkened room. Figure 8 shows the measured EOTFs. The solid lines denote fits of the function L:

$$L = a + (b + kV)^{\gamma}$$

to the measured data. Here, L denotes luminance, and V denotes 8-bit pixel value; the measured parameters are shown in the legend of Fig. 8 for each display.

The contrast thresholds were measured by following the same procedures as used in Alam et al.;[39] a three-alternative forced-choice procedure guided by a Quest staircase with a fixed 48 trials, a 10-ms time-limit per stimuli presentation, and audio feedback (see Ref. 39 for additional details). The RMS contrast of the distortions as defined in Sec. 3.1 and as used in Ref. 37 was used as the contrast measure. The mean luminance of the solid background upon which the three stimuli choices were placed was fixed at 2 cd/m^2, which is darker than used in Ref. 39, but required in order to obtain the same mean luminance across all display/brightness-setting/lighting variations. The viewing distance was adjusted for each display such that the image always subtended 4×4 deg of visual angle. Three trained male adults with normal or corrected-to-normal vision (YZ, YY, and DC, the three authors of this section) served as subjects in the experiment.

Figure 9 shows the resulting contrast detection thresholds. We performed a two-way, repeated-measures analysis of variance (ANOVA) with contrast threshold (in dB) as the dependent variable, and with display (Display++, I-O Data, iPad) and image (Shroom, SunsetColor) as the within-subject (repeated) factors. For this analysis, we used the thresholds averaged across trials from each subject, resulting in 18 average thresholds (3 displays $\times$ 2 images $\times$ 3 subjects). The analysis revealed that there was no significant main effect of display on threshold ($F_{2,4} = 0.68$, $p = 0.557$). There was also no significant main effect of image on threshold ($F_{1,2} = 9.71$, $p = 0.089$). There was a significant interaction effect ($F_{2,4} = 8.12$, $p = 0.039$), indicating that the

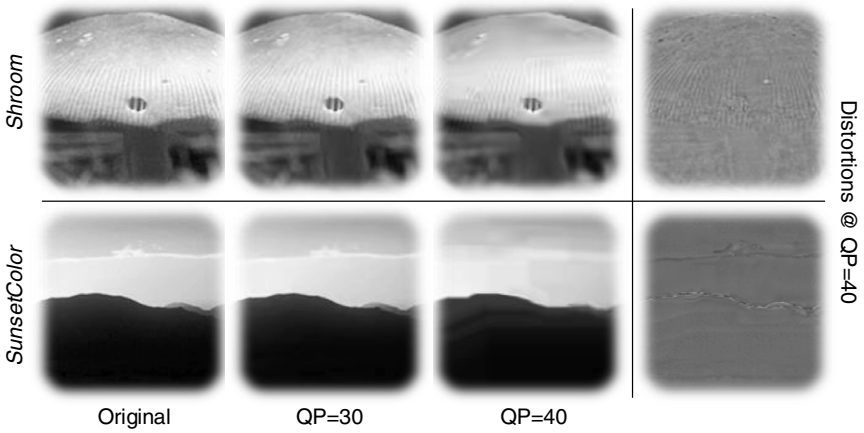


Fig. 7 Stimuli used in the study—original and HEVC-compressed image segments from the CSIQ image quality and masking databases.[1]

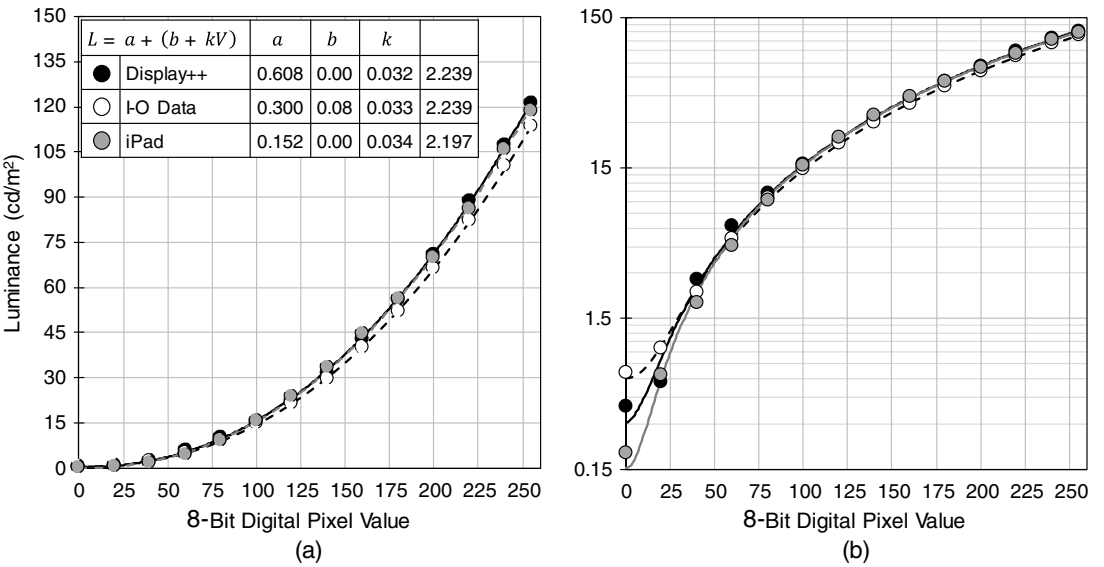


Fig. 8 EOTFs of the three displays on (a) linear and (b) logarithmic luminance scales.

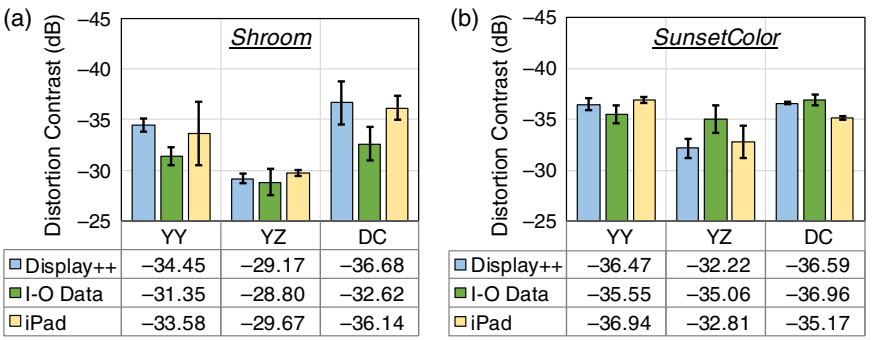


Fig. 9 Contrast detection thresholds on different displays. Each error bar denotes ±1 standard deviation of the respective mean. Note that the vertical axis is reversed, and thus, taller bars represent lower thresholds. (a) Shroom and (b) SunsetColor.

display has a different effect on the threshold, depending on the image.

Figure 10 shows plots of the marginal mean thresholds for each image (horizontal axis), with separate lines representing the different monitors. As shown in this figure, the fact that the three lines are not parallel indicates the interaction, which results from the I-O data monitor. Specifically, for Shroom,

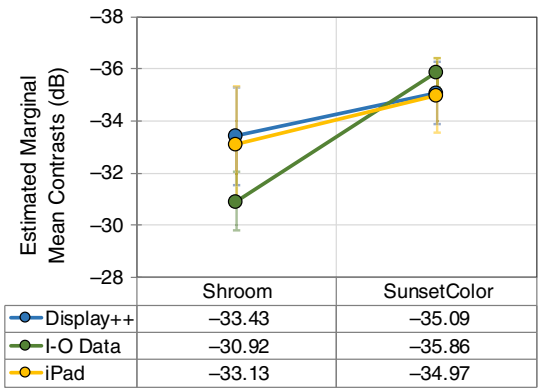


Fig. 10 Profile plots of the marginal mean thresholds showing the interaction between display and image. Each error bar denotes ±1 standard error of the respective mean.

the I-O data display yielded the highest average threshold (−30.7 dB), whereas the CRS and iPad displays yielded lower thresholds (−33.5 dB and −33.0 dB). However, for SunsetColor, the I-O data display yielded the lowest average threshold (−35.8 dB), whereas the Display++ and iPad displays yielded higher thresholds (−35.1 dB and −34.7 dB). However, Bonferroni-corrected posthoc analyses on the results for each separate image showed no significant simple effect of display on threshold ($F_{2,4} = 5.71$, $p = 0.067$ for Shroom; $F_{2,4} = 0.47$, $p = 0.657$ for SunsetColor).

Although only three subjects were tested, some preliminary comparisons can be made between the variations in thresholds due to display versus due to subjects. For image Shroom, the standard deviation across displays was ∼1.5 dB (averaged across subjects), whereas the standard deviation across subjects was ∼3 dB (averaged across displays). For image SunsetColor, the standard deviation across displays was ∼2 dB (averaged across subjects), whereas the standard deviation across subjects was ∼1 dB (averaged across displays).

Although only two images were tested, these results would seem to suggest that thresholds measured in the laboratory setting (by using a specialized display such as Display++, and to a lesser extent, a consumer-grade monitor) can yield thresholds, which are valid when the content is

viewed on an iPad. Similarly, for the stimuli used in this study, thresholds can be measured directly on a mobile display. Interesting, the profiles in Fig. 10 would seem to suggest that the iPad did a better job at yielding marginal means similar to those obtained on the Display++ monitor as compared to the I-O data monitor. These suggestions, of course, assume that viewing conditions and EOTFs remain similar. A further discussion on differences between the three displays is provided in Sec. 3.4.

3.3 *Effect of Display Setting*

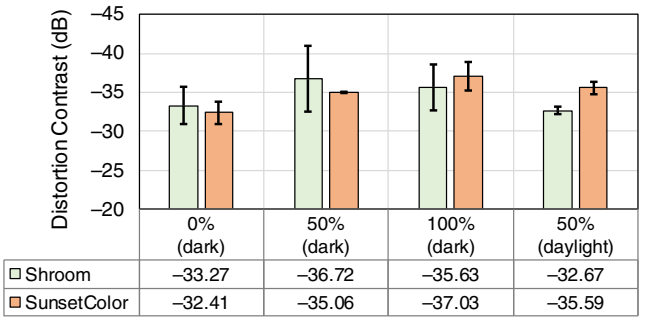
The ability to measure thresholds directly on a widely used mobile device, such as the iPad, enables the possibility of measuring thresholds via crowdsourcing. However, subjects might erroneously adjust the iOS "brightness" setting, thereby affecting the EOTF and ultimately affecting the thresholds. Similarly, subjects might mistakenly perform the experiment in a nondarkened room, thereby affecting the thresholds.

Thus, in a follow-up pilot experiment, we measured thresholds on the iPad under three iOS "brightness" settings: 0%, 50%, and 100%; and at 50% in a room lit by daylight (as opposed to a darkened room). The stimuli and procedures were identical to the previous experiment. Only the third author of this section (D.C.) participated in this pilot experiment.

Figure 11 shows the EOTFs of the iPad under these different settings. Observe that the iOS "brightness" setting primarily affects the slope on a linear luminance scale (vertical offset on a logarithmic scale); this is captured in the fits by the parameter k. However, the "brightness" setting also has a small effect on the minimum brightness (parameter a). Similarly, changing the room illumination from a darkened room to a room lit by daylight primarily raises the low end of the curve with negligible effects for larger luminances; this is captured by changes to parameters and a and b.

The resulting thresholds are shown in Fig. 12. To evaluate the effect of the "brightness" setting, we performed a two-way ANOVA with contrast threshold (in dB) as the dependent variable, and with "brightness" setting (0%, 50%, 100%) and image (Shroom and SunsetColor) as the factors. For this analysis, we used the per-trial data from the single subject, resulting in 24 thresholds (3 "brightness" settings × 2 images × 4 trials). The analysis revealed that there was a significant main effect of "brightness" on threshold ($F_{2,18} = 4.84,\ p = 0.021$). There was no significant main effect of image on threshold ($F_{1,18} = 0.01,\ p = 0.914$), nor was there a significant interaction between "brightness" and image ($F_{2,18} = 2.17,\ p = 0.143$).

A Bonferroni-corrected posthoc analysis revealed a significant difference between the 0% and 50% "brightness" ($p = 0.035$). As shown in Fig. 12, lowering the "brightness" to 0% raised the thresholds for both images (+3.45 dB and +2.65 dB for Shroom and SunsetColor, respectively). We suspect that this threshold elevation is attributable to a reduction in contrast sensitivity due to noise masking (increased variance of the internal decision variable): The reduced luminance range of the display made it difficult to see both the distortions and image.[40] The average luminance of the images under this setting was 1.4 and 1.2 cd/m^2 for Shroom and SunsetColor, respectively, presented against a fixed 2 cd/m^2 background. As recently measured by Kim et al.,[41] and as modeled by both the Daly CSF model[42] and the



	0% (dark)	50% (dark)	100% (dark)	50% (daylight)
Shroom	−33.27	−36.72	−35.63	−32.67
SunsetColor	−32.41	−35.06	−37.03	−35.59

Fig. 12 Contrast detection thresholds on the iPad under different settings/room illuminations. Each error bar denotes ±1 standard deviation of the respective mean. Note that the vertical axis is reversed, and thus taller bars represent lower thresholds.

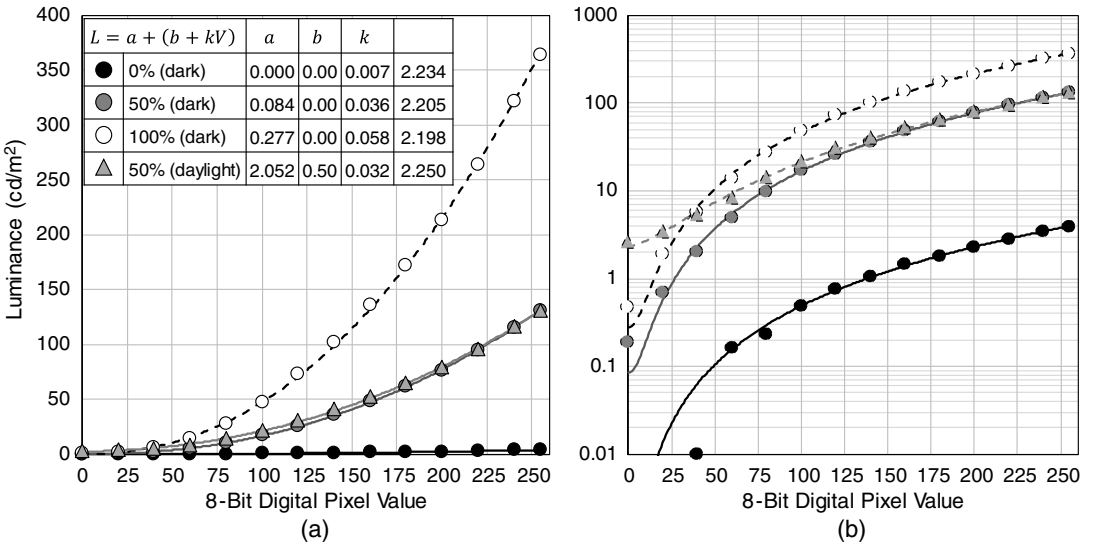


Fig. 11 EOTFs of the iPad with different iOS "brightness" settings and in darkened versus daylight room settings on (a) linear and (b) logarithmic luminance scales.

Barten CSF model,[43] for spatial frequencies greater than 1 cycle/deg, contrast sensitivity drops markedly at low luminance levels (2 cd/m^2), whereas as measured in Kim et al.,[41] sensitivity is nearly equal for the 20 and 150 cd/m^2. Although the reduced visibility of the image may very well have reduced the amount of luminance and contrast masking, for correlated distortions, oftentimes the mask (image) and target (distortions) are visually captured as a single percept, and therefore, subjects often look for mangled content rather than for separate distortions.[44,45] The reduction to 0% possibly inhibited the ability to see the mangled features, potentially giving rise to greater noise (internal) masking.[9]

The posthoc analysis also revealed that there was no significant difference between the 0% and 100% settings ($p = 0.058$) nor between the 50% and 100% settings ($p = 1.0$). This lack of change between 0% and 100% might be attributable to an increase in sensitivity due to a reduction in noise masking, countered by a decrease in sensitivity due to luminance masking. Furthermore, the similarities between the thresholds for the 50% and 100% conditions are as expected. At 50% "brightness," the images had average luminances of 45 and 39 cd/m^2, for Shroom and SunsetColor, respectively. At 100% "brightness," the averages luminances were 127 and 110 cd/m^2, for Shroom and SunsetColor, respectively. Assuming the adapting luminance was within the 20 to 150 cd/m^2 range, the CSFs for these two "brightness" settings should be the same. Furthermore, the RMS contrasts of the images were nearly identical under both the 50% and 100% settings (0.9 for Shroom and 1.1 for SunsetColor), suggesting the same amount of contrast masking under both settings.

Finally, to compare the 50% "brightness" setting under the two room-lighting conditions, we performed a two-way ANOVA with contrast threshold (in dB) as the dependent variable, and with lighting (daylight, dark) and image (Shroom and SunsetColor) as the factors. For this analysis, we used the per-trial data from the single subject, resulting in 12 thresholds (2 lighting conditions × 2 images × 4 trials). The analysis revealed no significant main effect of lighting ($F_{1,12} = 2.75$, $p = 0.123$), no significant main effect of image ($F_{1,12} = 0.39$, $p = 0.546$), and no significant interaction between the two ($F_{1,12} = 4.25$, $p = 0.062$).

Based on the EOTFs for the 50% dark and 50% daylight conditions (see Fig. 11), the main difference is in the lower-pixel-value range, where the EOTF for the daylight condition is elevated. This relative inability to produce low luminances should result in images of lower contrast, which was indeed the case (0.9 versus 0.8 for Shroom and 1.1 versus 1.0 for SunsetColor). Lower mask contrasts would suggest lower contrast masking, and thus reduced thresholds. Indeed, a slight (though not significant) change of −0.5 dB was observed in the thresholds for image SunsetColor, which contains very dark mountains in the lower half of the image. For image Shroom, on the other hand, the key locations for detecting the distortions were: (1) mangling of the very low-contrast ribs in the mushroom's cap and (2) blurring of the low-contrast texture in the top of the mushroom's cap. Because these areas serve as cues, any reduction in the contrasts of these areas would inhibit the ability to detect these mangled features and should thus raise thresholds. Indeed, for Shroom, a (albeit not significant) change of +3.0 dB was observed. A follow-up study with more images and subjects could provide more insights.

3.4 *Discussion and Summary*

A longstanding unknown in regard to quantifying visual losslessness in compressed images and videos is the applicability and reliability of such measurements, especially in regard to mobile displays. In this preliminary work, we have shown that contrast detection thresholds for HEVC distortions in 8-bit images can be similar when measured (via a forced-choice procedure) on an iPad Air 2 as compared to when measured on desktop and laboratory displays if the EOTFs and ambient lighting conditions are controlled, and if luminance contrast (e.g., RMS contrast) is used to quantify the distortions. Some significant differences in thresholds between the displays were observed; however, whether a particular display would raise or lower thresholds was found to be image-dependent. For the limited conditions tested in this study, the variation in thresholds across monitors was roughly as large as the variation in the thresholds across subjects.

In regard to the iPad, for the limited stimuli used in this study, the thresholds were surprisingly robust to reasonable variations in the iOS "brightness" setting and/or room illumination. Lowering the "brightness" setting from 50% to 0% raised thresholds, possibly attributable to a reduction of contrast sensitivity due to the low stimulus luminances. Raising the "brightness" setting from 50% to 100% did not significantly change the thresholds, again possibly attributable to the no change in contrast sensitivity and no overall change in the masks' RMS contrasts (due to the 100% "brightness" setting's increase of both high and low luminances). At 50% "brightness," thresholds measured in a dark room were not significantly different from those measured in a daylit room.

It is important to note that our findings do not imply that the display used in psychophysical studies does not matter. Different types of studies may certainly require bit-depths, ranges, and other properties, which are not possible on mobile or consumer-level desktop displays. Likewise, our findings do not imply that one can forgo characterization/calibration of the display and instead measure distortions in terms of code values (pixel values and quantization step sizes).

Given that the three displays were approximately matched in EOTFs, it is reasonable to expect that the thresholds measured on the three displays to show no significant differences. However, LCD displays can differ in other aspects, which were not measured in the current study, nor for which corrections were made. In particular, the displays likely differed in terms of the dependence of luminance on viewing angle (a problem with all LCDs), the spatial variability of luminance, and the extent to which the luminance at one position of the screen is influenced by the luminances at other positions. As reported in Ref. 46, although the Display++ display is robust to small viewing-angle changes (up to at least 15 deg), and although luminances remain largely uninfluenced by simultaneous luminances at other positions (<1 cd/m^2 change), the spatial uniformity was reported to vary up 18%. These aspects have not been measured for the I-O data or iPad displays used in the present study. It remains unclear whether these aspects affected our results.

Another difference between the three displays was the viewing distance, which was adjusted for each display to ensure that the degrees of visual angle subtended by the stimuli on all displays were the same, thus ensuring the same spatial frequencies of the targets and masks on all displays. For the Display++ monitor (the largest physical display), a viewing distance of 86 cm was used such that each image subtended ~4 deg of visual angle. For the iPad, a viewing distance of 22 cm was used to achieve the same 4 deg of visual angle. This large change in viewing distance necessitated different levels of accommodation, convergence, and pupil constriction from the subjects,[47] which could have influenced the perception of the target, mask, and/or their interaction. On the other hand, the large change in viewing distance also seemed to influence the quality-of-experience: All subjects reported that the experience on the iPad seemed more immersive, and thus gave the impression of greater detectability. Again, it remains unclear whether these aspects affected our results.

One other difference between the three displays was the coatings of the screens. The I-O data display had a matte antireflective coating, whereas the iPad and Display++ had glossy coatings. The iPad's glossy coating is reported to be antireflective, yet it was much more reflective than the matte I-O data display's screen. The properties of the glossy coating of the Display++ are unknown. Although we did not measure the MTF of each monitor, subjects reported that the images appeared slightly sharper on the iPad and Display++ screens as compared to the I-O data screen. In addition, although reflections were largely eliminated via the use of a black curtain, some reflections were particularly present on the iPad due to light from the display reflecting off of the subject's face and back onto the screen as a result of the close viewing distance. Again, it remains unclear whether these aspects affected our results.

Another important unanswered question is how thresholds for detecting compression distortions might change when using naïve versus expert subjects. In the present study, all three subjects were familiar with the purpose of the experiment, and all subjects performed a several practice sessions on each image before the start of each experimental session. Two of the subjects (Y.Z. and D.M.C.) had extensive prior experience with forced-choice detection experiments. We do not expect knowledge of the purpose of the experiment to have a marked effect on the thresholds. One of the advantages of forced-choice detection experiments as compared to subjective rating experiments is that the former is much more objective because it requires a simple choice rather than an opinion score, and response bias can be controlled by randomization of target choice location or interval. Furthermore, with a 3AFC procedure as opposed to a 2AFC procedure, the subjects do not need to know what changes to look for, they need only choose the odd one out. Knowing that the experiment is testing distortion visibility could inform the subject of where the distortions are likely to appear (e.g., those familiar with compression could look for blurring and/or ringing around edges), however, these locations are quickly learned within the first few practice sessions. Thus, training is important (granting familiarity with the mask and how it might change; a signal-known-exactly condition), but specific knowledge of the purpose of the experiment would likely not affect the results.

4 Subjective Assessment and the Criteria for Visually Lossless Compression

Objective metrics of image quality have the advantage of repeatability and are suitable for automatic assessment and monitoring of image quality. Such metrics are in high demand, given the increasing requirements for real-time image compression needed to deal with the bandwidth requirements of high-resolution image transmission.[22,48] However, it is clear that while subjective testing is labor intensive and costly, it remains the only reliable means of evaluating the impact of image compression and the visibility of artifact. As described in Secs. 1–3, a wide range of qualitative methodologies are available: both threshold and suprathreshold methods have been widely employed to assess image quality (and the success of compression algorithms). Forced-choice threshold methods are often used to establish if a compression algorithm is visually lossless as they are sensitive measures of the visibility that are less impacted by bias and amenable to statistical analysis.

In 2015, International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC) jointly published evaluation protocols based on forced-choice procedures that could be used to evaluate images across display platforms. Their protocol[49] describes two variants: one normative (Annex A) and the other (Annex B) based on a flicker paradigm proposed by Hoffman and Stolitzka.[19] In the normative approach, the original image is presented as a reference and, in another part of the display, the observer is presented both the reference image and the processed image SBS (randomly ordered) and required to choose which of these pair matches the reference image. This is a classical forced-choice procedure intended to measure sensitivity to artifacts in the processed image.

However, there are several issues that suggest that other procedures might be more appropriate. First, while there may be salient artifacts in the image, observers may not know where to look. Until attention is brought upon these areas, the literature suggests that even large changes in the image are often not seen.[50,51] This issue is partly addressed in the ISO/IEC protocol because instead of using large full screen images, the stimuli are crops of a predefined size that can be selected to include potentially problematic subregions. Techniques to highlight the changes could make for more sensitive and efficient detection (see Sec. 5 below). Other important use cases involve dynamic detection of visibility. For instance, in video or interactive content, frame-to-frame differences in quality might be noticeable. The second variant described in ISO/IEC 29170-2:2015 (Annex B)[49] uses direct temporal comparison in a flicker/toggle paradigm (see also Sec. 5). In this procedure, two image sequences are shown SBS, as illustrated in Fig. 13. On one side of the display, the reference image is shown alternating with the processed image (at a rate of 5 Hz), while on the other side, the reference image alternates with itself (i.e., does not change).

In this procedure, the reference is presented sequentially in the same location as the processed image; the image differences should be extremely salient due to sensitive motion and change detectors in the visual system. The flicker

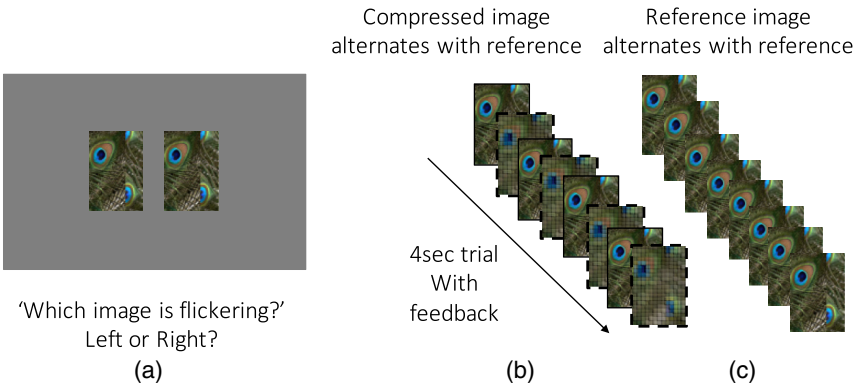


Fig. 13 Illustration of the ISO/IEC 29170-2:2015 flicker protocol (Annex B).[49] (a) The observer's view of the stimuli and (b) illustration of the image alternation (5 Hz). The reference location is randomized on each trial, viewers have 4 s to view the image sequences and are given feedback.

paradigm is also relevant to cases where transient image artifacts may occur such as video and interactive content. Extensions of this approach have recently been proposed to assess the quality of compression for HDR imagery and for comparison of still, panning images, and moving images.[52]

In a large-scale, trial ($N = 120$) conducted at York University, we implemented the ISO/IEC 29170-2:2015 flicker protocol[49] to assess the qualitative effectiveness of the Video Electronics Standards Association (VESA) display stream compression standard (DSC1.2) using a wide range of image content, including known challenges to the algorithm (see Fig. 14).

As specified by the ISO/IEC protocol, in addition to the test conditions of interest, a number of obviously degraded control conditions were evaluated. These "control" conditions provided encouragement to participants and who otherwise were performing at threshold most of the time. In addition, performance on these trials was used to exclude observers who were not paying attention; observer data were only included if an individual scored ≥95% on control trials. Each condition was tested multiple times to arrive at a detection probability for each stimulus condition. For each observer, and each condition, the proportion correct detection was calculated. The ISO/IEC standard recommends reporting of summary graphs in the format shown in Fig. 15. Here, the mean proportion correct is plotted across observers with ±1 standard deviation and symbols indicating the best and worst performing observers (downward and upward oriented triangles, respectively).

As discussed above, the criterion used to define visually lossless is critical and is under debate. Given that the ISO/IEC 29170-2:2015[49] standard is based upon detection, a detection threshold approach is used. Following psychophysical convention, a 75% correct criterion (midway between guessing and perfect responses in the two-alternative task) is recommended. Specifically, the standard proposes that lossless performance occurs when no observer detects the compressed reference on greater than 75% of the trials (although the standard allows for modification of the criterion). In our study, the large majority of test conditions met these strict criteria for visually lossless (see Fig. 15). However, in some instances, this criterion may result miscategorization of reference conditions as lossy. The results shown in Fig. 16 illustrate this phenomenon. In this graph, under all compression conditions, average observer performance is clearly at chance (50% for this

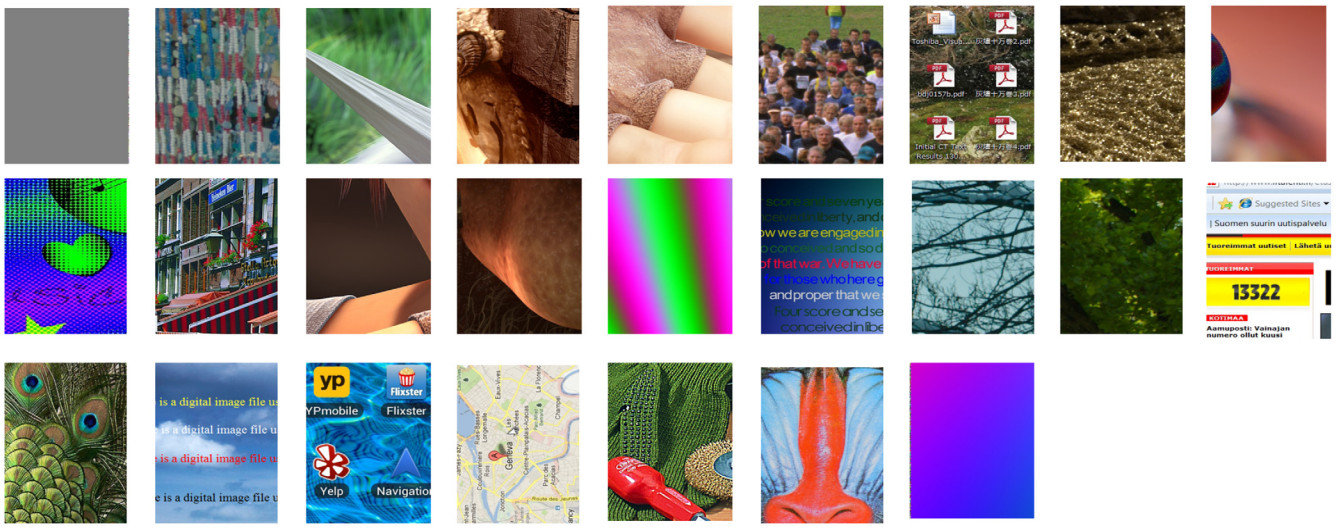


Fig. 14 Thumbnails of images used in the assessment of VESA DSC1.2. A wide range of compression parameters were applied to each image (chroma subsampling, lines per slice, bits per channel).

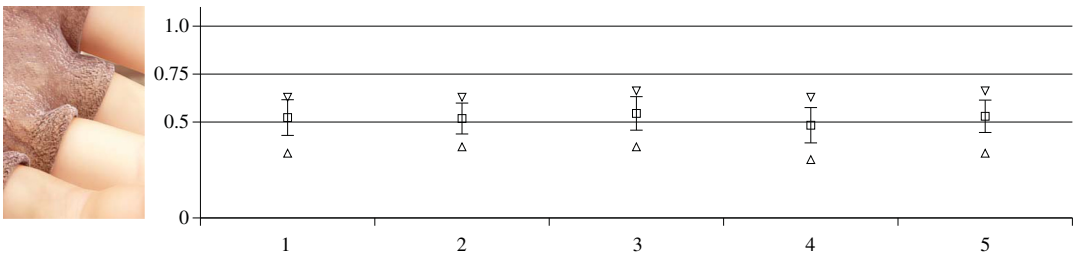


Fig. 15 The graph shows the proportion correct for a given image (thumbnail to left) under different compression conditions (coded as numbers 1 to 5). Open squares represent the group average, error bars indicate ±1 standard deviation, and downward and upward triangle symbols indicate the best and worst performance. Each dataset represents a different level of compression.

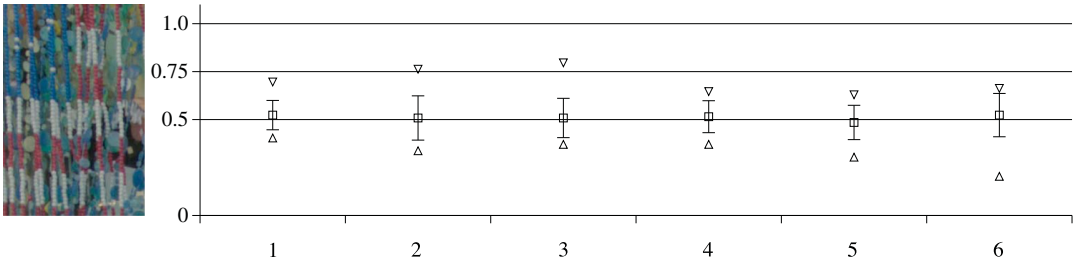


Fig. 16 As shown in Fig. 15, results are shown for here for six compression conditions for the image shown in the inset. Open squares represent the group average, error bars indicate ±1 standard deviation, and downward and upward triangle symbols indicate the best and worst performance.

two-alternative task); however, in condition 3, one observer detected the flicker on more than 75% of the trials.

It is clear that this criterion places considerable emphasis on potential outliers in the dataset. In their original implementation of the flicker protocol, Hoffman and Stolitzka[19] identified and selectively tested a set of 19 (out of 35) highly sensitive observers in their dataset. They suggest that given the potential impact of such observers that the criterion for lossless could be increased to 93%, but just for these sensitive individuals. However, this approach introduces a potential bias to the test protocol: it is left to the experimenter to define the sensitive observers, who will be held to a different standard. Another approach would be to consider the results of all observers, but to adopt a visually lossless criterion based on their average performance and the associated standard deviation (for example, using the estimated 95th percentile rather than the sample maximum). Statistical techniques based on the variance could be used to identify highly sensitive observers or outliers, and, if appropriate for the use-case, remove them from the dataset.[53]

Another factor that contributes to the sensitivity of the ISO/IEC protocol is the extent to which practice on a limited image set can, over time, contribute to the creation of highly trained observers. Such observers could learn to attend to specific image regions, and as a result be able to better detect artifact-related flicker. As noted in previous sections, SBS presentation allows observers to directly compare reference and original image regions, improving detection rates. Furthermore, in the paradigm implemented here, cropped rather than full screen, image regions are viewed. At the recommended viewing distance, each image is within highly sensitive foveal vision, which further enhances the probability of detecting the flicker created by alternating the reference and original images. These factors, combined with the sensitivity of the human visual system to spatiotemporal variation within this range, will draw attention to compression-related distortions. Over time and trials viewers may become attuned to specific regions and artifacts that would otherwise remain undetected. These training effects could be reduced by using a large pool of observers on a limited set of conditions, but reduction of test content will in turn reduce the generalizability of the evaluation.

The results of our evaluation of VESA DSC1.2 using the ISO/IEC 29170-2015 flicker protocol show that this forced-choice paradigm is a highly effective means of evaluating sensitivity to image differences.[49] The design of this test protocol and the visually lossless criterion applied is extremely sensitive, and in particular, emphasizes the most sensitive viewers. It is arguable that this protocol is too sensitive, and that the results of the SBS flicker task highlight artifacts that would not ever be visible under "normal" viewing conditions. As outlined elsewhere in this paper, there are other candidate approaches to the assessment of image quality. We argue that the appropriate methodology depends on the objectives and the use case. For example, if the goal is to conservatively evaluate the possibility that a compression artifact might be visible under any situation, then the flicker paradigm is a viable approach as it highlights differences between images regardless of whether they are noticeable in the absence of a reference.

5 Usage Perspectives on Visually Lossless and Lossy Quality and Assessment

The display industry today struggles to distinguish between visually lossy and lossless encoding, as visual experiences are determined by various factors, such as usages, form factor, and content. The debate becomes even further entangled when one tries to define and quantify visually lossless based

on empirical measurements, when the chosen measurement protocol and stimuli can critically impact the outcomes.

As outlined in the previous section, Wilcox et al., VESA has presently adopted a testing protocol and procedure for evaluating visually lossless encoding. From a practical perspective, this is only one particular way to investigate whether a process or algorithm can potentially produce a result that is truly visually lossless. Such a specific approach was reasonable when research first started on this topic, as the problem space in which we needed to investigate was clear, and the usages and context for the definition could be clearly carved out. However, the increasingly complex ecosystem in electronic displays warrants a reinvestigation and redefinition.

Innovations in testing methods and their underlying theories open up the possibility of using different techniques, such as gaze tracking, to potentially augment and improve existing methodologies by emphasizing the visual functions in specific usages.

The proposed research emphasizes the need of refining the definition of visually lossless and delineating the testing procedure based on the usages and viewing context.

Classical visually lossless compression of display content is defined as the loss of image quality induced by the compression algorithm that cannot be perceived by a user.[54] This definition by its nature is very vague, as the quality of perceived image can be affected by individual visual characteristics, viewing conditions, and display apparatus. Thus, there is no unified position on what the definition of visually lossless truly is. Conversely, in the display industry, it has been deemed that each company, organization, or institution can set thresholds based on their products and desired experiential delivery. What this means at its core is that the statistical requirements for user studies can be set at different levels and different means of detection. It is the suggestion of the authors that we use a more traditional definition of visually lossless with tighter statistical constraints. To develop a commonly accepted definition of visually lossless, it becomes necessary to provide a unified principle of assessing these outcomes that takes into considerations the above factors, using a 5% criterion for loss detection.

In 2014, VESA published a new standard that uses visually lossless image compression to increase the rate of data transmission carried by a display interface data rate, thus saving power while maintaining viewer's experiences.[49] As part of the new standard, visually lossless with regard to visual psychophysics requires that any content distortion caused by compression be below the threshold of conscious detection. Therefore, visually lossless compression should be defined as providing for as any loss of data due to file compression that is not detectable to a "typical" user on a "typical" display under "typical" viewing conditions. This has been implemented as a SBS comparison of control (original to original) and target (original to compressed) images alternating (flickering) at 2.5 Hz. A user is asked to discern the target image from the control image within 30 s of viewing. Outcomes of such method can be affected by factors, such as viewing distance, screen luminance, pixel density, viewer's visual acuity, etc.

Not surprisingly, the use of the ambiguous term "typical" has been proven problematic when it comes to verification testing. The definition and assessing methods of visually lossless compression have been guided by the ecology of display industry. When the first discussions of visually lossless compression began, the ecosystem of media was relatively simple as the usage did not vary and the form factor uniform. With the pervasiveness of media content in the world today, the fundamental definition of visually lossless or "lossy" needs to be examined at a contextual level. When one moves from consumption of content on a phone to a large format TV or to new VR and AR environments, the nature of the changed viewing environments dictates that the definition of visually lossless will need to be expanded from less immersive traditional visual environments to account for more fully immersive nontraditional environments.

With the adoption and deployment of VR devices, the issue becomes even more complex due to the artificial binocular delivery of the stimuli to the visual system. With it, vision is not just the result of stimulus-derived representations, but also that of interaction between visuomotor processes. Many feel that this will cause the visual experiences to be different under different usages, and the definition of lossless vision to be bifurcated between monocular and binocular devices. These would suggest a need of different methodologies for testing visually lossless compression according to involved visual imagery as well as underlying visuomotor functions.

In the previous section, it is mentioned that the definition of visually lossless has been left to the manufacturer of the devices, which opens an interesting competitive angle in determining, who has the best performance and who will or will not claim performance based on any given standard implementation. The authors feel that as the definition of visual experiences has evolved, there is a large gray area that needs to be explored around performance in order to have contextually correct definitions. There is still room for debate around what could be standardized versus not and whether there are more generalized testing methods based on not specific devices but utilized human visuomotor processes.

Copious research has demonstrated that human vision is not a replica of the visual world, but an outcome of interactive visuomotor processes.[55] Scientists have commonly identified two types of human vision: featural and spatial.[56,57] Features, such as shape, color, and complex object categories, are encoded in the ventral aspect of human cerebral brain,[58] whereas different spatial representations, such as retina- and body-centric ones are encoded in the dorsal aspect. Featural perception dominates the conscious vision and is generated as much by bottom-up visual stimuli as by top-down insertion and creation of visual imagery. The spatial information is utilized by the brain for forming perception and guiding complex actions but is not directly accessible to the perception.[59,60] When viewers process displayed visual content, much information is utilized by the brain but not consciously identified. Visual attention serves to combine the two types of vision and makes some aspect of it available to visual consciousness.[59,61] Hence, functionally lossless vision should be defined as providing for unimpeded visuomotor processes in maintaining such interactive representations of visual world. The purpose of the visual tasks, the predominant context (static or in motion), and the level of focused attention determine the threshold of lossy vision.

We suggest it is useful but insufficient to simply compare compressed images to uncompressed ones for determining

lossless vision. Such outcomes need to be obtained in a realistic task consistent with the form factor and usage, in which the task goals determine what should be attended and at which conscious level. Visually lossy should then be defined as either the detection of visual degradation or impeded execution of visual tasks.

To discern lossy vision, we propose to utilize a well-recognized method of gaze-contingent image degradation (GCID). GCID is achieved by switching between an original image and degraded image during eye fixations or saccadic eye movements. The perception of stable visual world is the result of visuomotor integration across eye fixations, where relevant visual percepts are maintained and unrelated features discarded. Instead of comparing SBS flickering images, in GCID, the original image was switched to the compressed ones during the critical part of eye fixation, e.g., 150 ms after fixation onset. The visuomotor performance and eye behaviors during selected eye fixations with control (original-to-original) and target (original to compressed) images are compared. This allows the effect of difference in visual imagery to be separated from artificial stimulation, such as by persistent image flickering. The GCID does not require the viewer to direct the foveal vision toward a specific area in juxtaposed images and yet the image switch is always available to the foveal vision, the viewing behavior is minimally altered, and the allocation of visual attention is guided by the content of the image. Furthermore, as the eye movements in GCID are guided by internal processes to search for and utilize necessary visual information, impeded vision is readily identified with a change in eye movement pattern, as documented in previous studies.[45] Lossy vision is present when the visuomotor process is slowed or altered, as determined by increased eye fixation duration, reduced saccade length, and higher frequency refixation when the original image was substituted with the compressed image during eye fixation.

Our preliminary data have demonstrated the effectiveness of such a paradigm. In the study, the subject was asked to indicate the location of image degradation with a mouse clicking at the end of 15 s viewing and was not informed of the image switch taking place during eye fixations. The data were obtained with a single switch between original and degraded (blurred) image at 100 ms after the onset of eye fixation; Fig. 17 shows that the location of degraded image (subtle blur) was detected at a level beyond chance (93%). In addition, the fixation duration (or latency of saccadic eye movements) within the blurred areas was significantly increased (250 to 269 ms) and saccade amplitude decreased (3.05 to 2.71 deg). Figure 18 shows that saccade initiation probability (as calculated from saccade hazard level[60]) was reduced when a blur was present, suggesting impeded visuomotor processing. Such eye movements and eye behaviors are tightly linked to the task on hand and the stimulus being processed. Therefore, this method can be useful to assess lossless images by measuring the proper baseline responses and changes in them caused by lossy images. An algorithm can be regarded as producing lossless image if the rate of detection and the parameters of eye movements are not significantly above change and deviating from the viewing of a static, unaltered image; an algorithm producing lossy images can be identified by the above-change detection of image degradation and altered eye

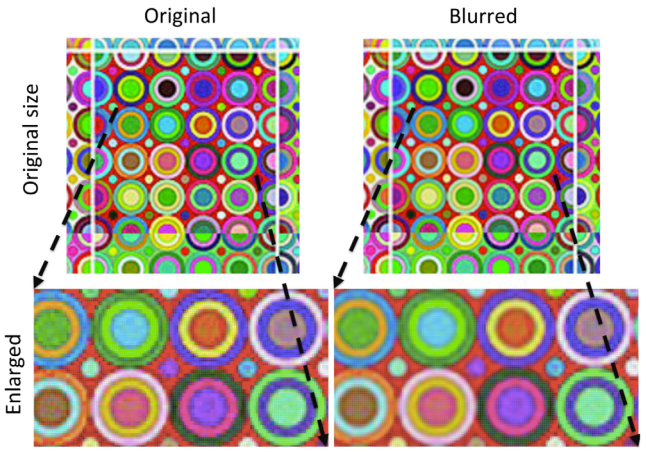


Fig. 17 Example original (left) and degraded images (right, subtle blur). Human subjects were asked to survey the image and reported the degraded area of the image. The degraded image was made available by switching from the original image to the blurred image at 100 ms after fixation onset. Human subjects were also to detect the blur in despite of the subtle change (93%).

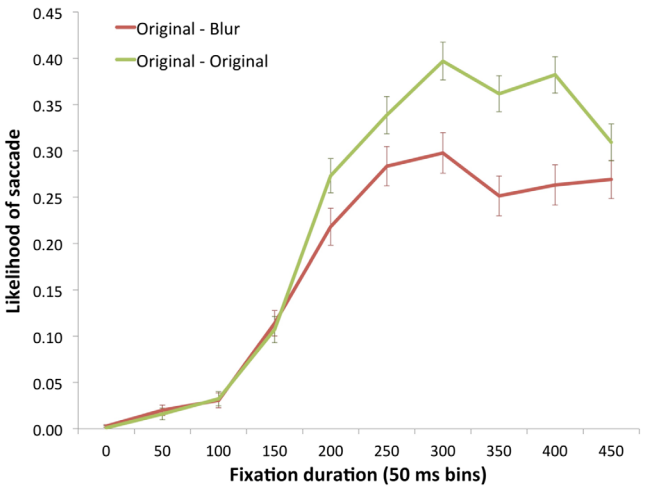


Fig. 18 Saccade probability calculated from the change in fixation duration using hazard function analysis. For fixations with degraded (blurred) images, a large proportion of saccades were delayed. The detection of blurred image (subtle blur, square symbols) has a latency of 200 ms (i.e., 100 ms after image switch), when the saccade probability began to be reduced compared to without image degradation (same image, triangle symbols).

movements. Such a paradigm can be utilized to evaluate lossy vision involving different ocular demands (e.g., performing a visual task at a close distance) or methods of image rendering (e.g., VR/AR displays).

Further empirical results based on the GCID paradigm have been published recently and have shown the great sensitivity of eye movement parameters to the appearance of degraded images regardless of whether the viewers were consciously aware of it; in contrast, the existence of constantly flickering images led to a significant rate of false alarm, likely due to heightened visual attention.[47] Thus, the presently adopted flickering paradigm could be unnecessarily strict and reject algorithms that would be perceived as lossless in many form factors and applications. Conversely,

the gaze-tracking method can permit greater lossless bandwidth compression where it is required.

To conclude, the authors propose that there is a need for at least two types of complementary methods in assessing lossless compressions: vision-for-perception and vision-for-action. These should be chosen based on the specific nature of display usage. For VR/AR and gaming displays, GCID and eye movement measures allow the rate of image and audio update to be optimal while permitting more forgiving compression algorithm. For home theater TV and high-resolution displays, detection of flickering image would allow better assessment of lossless algorithms when image degradation is more easily discerned. Therefore, the gaze-tracking paradigm can be a very useful tool for many display applications although it is harder to utilize due to its requirement of specialized equipment.

6 Conclusions

In this paper, we have attempted to bring together a variety of academic and industrial studies/use-cases that rely on the notion of visually lossless image quality.

In Sec. 2, Daly describes a variety of business perspectives on visual losslessness. The main conclusion from this section is that the appropriate method of video comparison (simultaneous versus sequential versus oscillation), the required level of display calibration, and the criteria (visually lossless versus visually lossy) depend on the business. Visually lossless criteria are relevant for mature businesses already delivering a high quality, whereas businesses for which visually lossy quality ratings are more relevant include newly developing businesses, developing products offering new features and conveniences, and businesses specializing in lower-cost products.

In Sec. 3, Zhang et al. have investigated visual losslessness in the context of HEVC compression of 8-bit images via a visual detection (of artifacts) experiment employing a three-alternative forced-choice paradigm. The main conclusion from this section is the suggestion that contrast detection thresholds for HEVC distortions can be similar on mobile, desktop, and laboratory displays if the EOTFs are similar. Thus, it may be possible to measure thresholds on one type of display, and still use those thresholds to predict whether distortions will be visible on another type of display.

In Sec. 4, Wilcox et al. have described the results of a study to assess the effectiveness of the VESA DSC 1.2 display stream compression standard using the ISO/IEC 29170-2015[49] flicker protocol (a forced-choice paradigm). The main conclusion from this section is that VESA's forced-choice paradigm is a highly effective means of evaluating sensitivity to image differences, and in particular, emphasizes the most sensitive viewers. If the goal is to conservatively evaluate the possibility that a compression artifact might be visible under any situation, then the flicker paradigm is a viable approach.

In Sec. 5, Colett et al. have described the need for at least two types of complementary methods in assessing lossless compression (vision-for-perception and vision-for-action), which should be chosen based on the specific nature of display usage. The authors feel that there is a need for at least two contrasting methodologies that focus on different display usage when evaluating visually lossless compressions. For usages focusing on the sense of immersion and authenticity, such as picture and movie viewing, the assessment of lossless image quality should require a realistic task for detecting visual anomaly, such as flickering between original and compressed images. SBS comparison of flickering images is artificial and can impose overly strict standards that cause issues in image rendering for display usages, such as VR and AR. For usages demanding active user interaction, a method utilizing eye tracking is useful for assessing visual quality for both conscious recognition and unconscious response guidance.

Together, the studies presented in this paper suggest that a single definition of visually lossless is not appropriate. Instead, a better goal might be to establish varying levels of visually lossless (similar in spirit to p-value used in statistical tests), which can be quantified in terms of the testing paradigm. For example, one could define visually lossless under the oscillation paradigm for vision-for-perception" as perhaps the most sensitive, whereas visually lossless under the sequential paradigm for vision-for-action" as perhaps the least sensitive. In most industry applications, the display parameters and viewing conditions are either implied from common usage or specified in standards documents. Actual specification is obviously the preferred approach for many of the reasons described in this article. Furthermore, still images would need a different level as compared to moving images in videos. By establishing such a list of varying levels of visually lossless, one could choose the most appropriate level depending on the application (target audience, imagery, business). For example, medical applications might choose the most sensitive level, whereas mobile video applications might choose the least sensitive level. Thus, rather than converging on a single definition of visually losslessness, industry can instead select the level (or propose a new level) that best suit their use cases and objectives. The expansion of display usages demands distinct assessment methods for image compression that is deployed across products and platforms.

Acknowledgments

RISE Acreo's work was funded by Knowledge Foundation (Grant No. 20160194, http://www.kks.se), which is hereby gratefully acknowledged. The work by Y. Zhang, Y. Yaacob, and D. M. Chandler was funded, in part, by JSPS KAKENHI Grant 17K00232 and Suzuki Foundation Grant D0D199E0E1420100 to D. M. Chandler.

References

1. G. T. Barnes and K. Lauro, "Image processing in digital radiography: basic concepts and applications," *J. Digital Imaging* **2**(3), 132–146 (1989).
2. A. B. Watson, "Receptive fields and visual representations in human vision, visual processing, and digital display," *Proc. SPIE* **1077**, 190–197 (1989).
3. J. A. Ferwerda and F. Pellacini, "Functional difference predictors (FDPs): measuring meaningful image differences," in *37th Asilomar Conf. on Signals, Systems & Computers*, pp. 1388–1392 (2003).
4. M. Tan et al., "The perception of lighting inconsistencies in composite outdoor scenes," *ACM Trans. Appl. Percept.* **12**(4), 1–18 (2015).
5. ITU-R, "Methodology for the subjective assessment of the quality of television pictures (ITU-R Rec. BT.500-13)," International Telecommunication Union, Radiocommunication Sector (2012).
6. R. Gross et al., "Generalizing GANs: a turing perspective," in *31st Conf. on Neural Information Processing Systems*, pp. 6317–6327 (2017).
7. ITU-T, "Subjective video quality assessment methods for multimedia applications (ITU-T Rec. P.910)," International Telecommunication Union, Telecommunication standardization sector (1999).

8. C. R. Carlson and R.W. Cohen, *Visibility of Displayed Information*, Office of Naval Research by RCA Laboratories, Princeton, New Jersey (1978).
9. R. Thiel et al., "Assessment of image quality in digital cinema using the motion quality ruler method," *SMPTE Motion Imaging J.* **116**(2–3), 61–73 (2007).
10. J. M. Hillis and D.H. Brainard, "Do common mechanisms of adaptation mediate color discrimination and appearance? Contrast adaptation," *J. Opt. Soc. Am. A* **24**(8), 2122–2133 (2007).
11. A. Harris, "The hidden side of *Silence of the Lambs'* most famous scene," *Slate's Culture Blog*, 2014, http://www.slate.com/blogs/browbeat/2014/10/15/silence_of_the_lambs_video_essay_from_tony_zhou_and_every_frame_a_painting.html (21 September 2018).
12. D. P. Darcy et al., "Physiological capture of augmented viewing states: objective measures of high-dynamic-range and wide-color-gamut viewing experiences," in *IS&T Int. Symp. on Electronic Imaging, Human Vision and Electronic Imaging*, San Francisco, California, p. HVEI126, Society for Imaging Science and Technology (2016).
13. P. Hanhart et al., "Subjective quality evaluation of high dynamic range video and display for future TV," *SMPTE Motion Imaging J.* **124**(4), 1–6 (2015).
14. R. A. Bradley and M.E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika* **39**(3/4), 324–345 (1952).
15. U. Neisser, *Cognitive Psychology*, Appleton-Century-Crofts, New York (1967).
16. A. O. Dick, "Iconic memory and its relation to perceptual processing and other memory mechanisms," *Percept. Psychophys.* **16**(3), 575–596 (1974).
17. G. Sperling, "The information available in brief visual presentations," *Psychol. Monogr. Gen. Appl.* **74**(11), 1–29 (1960).
18. S. Magnussen, "Low-level memory processes in vision," *Trends Neurosci.* **23**(6), 247–251 (2000).
19. D. M. Hoffman and D. Stolitzka, "A new standard method of subjective assessment of barely visible image artifacts and a new public database," *J. Soc. Inf. Disp.* **22**(12), 631–643 (2014).
20. J. Froehlich et al., "Content aware quantization: requantization of high dynamic range baseband signals based on visual masking by noise and texture," in *2016 IEEE Int. Conf. on Image Processing (ICIP)* (2016).
21. H. R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.* **15**(2), 430–444 (2006).
22. Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *37th Asilomar Conf. on Signals, Systems & Computers* (2003).
23. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
24. R. Mantiuk et al., "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graphics* **30**(4), 1–14 (2011).
25. M. Narwaria et al., "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).
26. M. Narwaria, M. Perreira Da Silva, and P. Le Callet, "HDR-VQM: an objective quality measure for high dynamic range video," *Signal Process. Image Commun.* **35**, 46–60 (2015).
27. ITU-R, "Reference electro-optical transfer function for flat panel displays used in HDTV studio production (ITU-R Rec. BT.1886)," International Telecommunication Union, Radiocommunication Sector, https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.1886-0-201103-I!!PDF-E.pdf (2011).
28. S. Miller, M. Nezamabadi, and S. Daly, "Perceptual signal coding for more efficient usage of bit codes," in *The 2012 Annual Technical Conf. & Exhibition* (2012).
29. C. A. Poynton, "Rehabilitation of gamma," *Proc. SPIE* **3299**, 232–249 (1998).
30. SMPTE, "ST 2086:2014 - SMPTE Standard - Mastering Display Color Volume Metadata Supporting High Luminance and Wide Color Gamut Images," *IEEE Spectrum* (2014).
31. SMTPE, "ST 2094-1:2016 - SMPTE Standard - Dynamic Metadata for Color Volume Transform – Core Components," *IEEE Spectrum* (2016).
32. S. Daly et al., "41.1: distinguished paper: viewer preferences for shadow, diffuse, specular, and emissive luminance limits of high dynamic range displays," in *SID Symp. Digest of Technical Papers*, pp. 563–566, Blackwell Publishing Ltd. (2013).
33. ITU-R, "Parameter values for the HDTV standards for production and international programme exchange (Rec. ITU-R BT.709-5)," International Telecommunication Union, Radiocommunication Sector, https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.709-6-201506-I!!PDF-E.pdf (2002).
34. ITU-R, "Parameter values for ultra-high definition television systems for production and international programme exchange (ITU-R BT.2020)," International Telecommunication Union, Radiocommunication Sector, https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.2020-2-201510-I!!PDF-E.pdf (2012).
35. Wikipedia, "W. Edwards Deming," 2018, https://en.wikipedia.org/wiki/W._Edwards_Deming (5 July 2018).
36. ITU-T, "High efficiency video coding (ITU-T Rec. H.265)," International Telecommunication Union, Telecommunication Standardization Sector, https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-H.265-201802-I!!PDF-E&type=items (2016).
37. M. M. Alam et al., "Relations between local and global perceptual image quality and visual masking," *Proc. SPIE* **9394**, 93940M (2015).
38. S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," *IEEE Trans. Broadcast.* **54**(3), 660–668 (2008).
39. M. M. Alam et al., "Local masking in natural images: a database and analysis," *J. Vision* **14**(8), 22–22 (2014).
40. A. B. Watson, R. Borthwick, and M. Taylor, "Image quality and entropy masking," *Proc. SPIE* **3016**, 1–11 (1997).
41. K. J. Kim, R. Mantiuk, and K. H. Lee, "Measurements of achromatic and chromatic contrast sensitivity functions for an extended range of adaptation luminance," *Proc. SPIE* **8651**, 86511A (2013).
42. S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 179–206, MIT Press, Cambridge (1993).
43. P. Barten, *Contrast Sensitivity of the Human Eye and its Effects on Image Quality*, SPIE Press, Bellingham, Washington (1999).
44. J. Nachmias and R.V. Sansbury, "Grating contrast: discrimination may be better than detection," *Vision Res.* **14**(10), 1039–1042 (1974).
45. D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," *Proc. SPIE* **6057**, 60570F (2006).
46. M. Ghodrati, A. P. Morris, and N. S. C. Price, "The (un)suitability of modern liquid crystal displays (LCDs) for vision research," *Front. Psychol.* **6**, 303 (2015).
47. S. R. Bharadwaj, J. Wang, and T.R. Candy, "Pupil responses to near visual demand during human visual development," *J. Vision* **11**(4), 6–6 (2011).
48. A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE* **101**(9), 2008–2024 (2013).
49. International Organization for Standardization and the International Electrotechnical Commission, "DIS 29170-2, Evaluation procedure for visually lossless coding," International Organization of Standards (2015).
50. A. Mack and I. Rock, "Inattentional blindness: perception without attention," in *Visual Attention*, R. D. Wright, Ed., pp. 55–76, Oxford University Press, New York (1998).
51. D. J. Simons, "Attentional capture and inattentional blindness," *Trends Cognit. Sci.* **4**(4), 147–155 (2000).
52. D. Stolitzka, P. Schelkens, and T. Bruylants, "New procedures to evaluate visually lossless compression for display systems," *Proc. SPIE* **10396**, 103960O (2017).
53. J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading (1977).
54. L. Karam, "Lossless image compression," in *The Essential Guide to Image Processing*, A. C. Bovik, Ed., pp. 385–417, Elsevier Academic Press, London (2009).
55. P. S. Churchland, V. S. Ramachandran, and T. J. Sejnowski, "A critique of pure vision," in *Large-Scale Neuronal Theories of the Brain*, C. Koch and J. L. David, Eds., p. 23, MIT Press, Cambridge, Massachusetts (1993).
56. U. Ansorge, W. Kunde, and M. Kiefer, "Unconscious vision and executive control: How unconscious processing and conscious action control interact," *Conscious. Cognit.* **27**, 268–287 (2014).
57. M. Goodale and D. Milner, *Sight Unseen: An Exploration of Conscious and Unconscious Vision*, Oxford University Press, Oxford (2013).
58. A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.* **12**(1), 97–136 (1980).
59. J. Najemnik and W.S. Geisler, "Optimal eye movement strategies in visual search," *Nature* **434**(7031), 387–391 (2005).
60. S. N. Yang, "Effects of gaze-contingent text changes on fixation duration in reading," *Vision Res.* **49**(23), 2843–2855 (2009).
61. C. Koch and N. Tsuchiya, "Attention and consciousness: two distinct brain processes," *Trends Cognit. Sci.* **11**(1), 16–22 (2007).

Robert S. Allison is a professor of electrical engineering and computer science at York University and a member of the Centre for Vision Research. His research investigates how we can reconstruct and navigate the three-dimensional world around us based on the two-dimensional images on the retinas. His work enables effective technology for advanced virtual reality and augmented reality and for the design of stereoscopic displays. He is the recipient of the Premier's Research Excellence Award in recognition of this work.

Kjell Brunnström is a senior scientist at RISE AB (Acreo), leading visual media quality and adjunct professor at Mid Sweden University. He is a cochair of the Video Quality Experts Group (VQEG). His research interests are in quality of experience (QoE) for video and display quality assessment (2D/3D, VR/AR, immersive). He is associate editor of the *Journal of Advances in Multimedia* and

has written more than hundred articles in international peer-reviewed scientific journals and conferences.

Damon M. Chandler received his BS degree in biomedical engineering from Johns Hopkins University, Baltimore, Maryland, in 1998, and the MEng, MS, and PhD degrees in electrical engineering from Cornell University, Ithaca, New York, in 2000, 2004, and 2005, respectively. From 2005 to 2006, he was a postdoctoral research associate with the Department of Psychology, Cornell University. From 2006 to 2015, he was on the faculty of the School of Electrical and Computer Engineering at Oklahoma State University, USA. He is currently an associate professor with the Department of Electrical and Electronic Engineering, Shizuoka University, Japan. His research interests include image processing, data compression, computational vision, natural scene statistics, and visual perception.

Hannah R. Colett is a human factors engineer and software developer at Intel. Her specialty is the development, design, and implementation of applications for collecting subjective metrics from users for competitive assessment. Currently, she works in the sales and marketing arm of Intel working on the development of methodologies for measuring, collecting, predicting, and benchmarking the user experience associated with various competitive technologies. She also works extensively with her alma mater Oregon State University to bridge academia and industry needs.

Philip J. Corriveau is a principal engineer and director of UX at Intel Corporation and adjunct professor at Pacific University. In January 2009, he was awarded a National Academy of Television Arts and Science, Technology & Engineering Emmy® Award for User Experience Research for the Standardization of the ATSC Digital System. He manages a multidisciplined team investigating user experience metrics for Intel products and technologies. Currently, he works in the sales and marketing arm of Intel.

Scott Daly is currently with Dolby Laboratories and is working on HDR, high-frame rate (HFR), VR imaging systems, and auditory-visual interactions with a focus on perceptual issues. He has degrees in electrical engineering and bioengineering. Previous accomplishments include key contributions to DICOM, a technical Emmy, the visible differences predictor (VDP), an Otto Schade award from SID, and coauthor of the cone-based perceptual quantizer nonlinearity (SMPTE 2084).

James Goel has been employed as a director of engineering and technical standards by Qualcomm since 2011. He has a BSc degree in applied science and electrical engineering from the University of Waterloo, Canada, 1992.

Juliana Y. Long holds a BA in history and anthropology from University of California, Irvine, and a MA in anthropology from California State University Fullerton. She also holds an MS degree in human resources management from Chapman University. She is a user experience researcher and human resources professional, currently working at Intel Corporation in the Human Resources group, conducting research on the organization's culture and doing program management for qualitative user studies and fieldwork.

Laurie M. Wilcox is a professor of psychology at York University, Toronto. She is a long-standing member of the Centre for Vision Research and of the graduate program in biology. In addition to fundamental research on stereoscopic depth perception, she collaborates with industry partners on applied projects related to 3-D cinema, image quality assessment, and electronic display systems. Her work is funded by several sources, including the Natural Sciences and Engineering Research Council of Canada.

Yusizwan M. Yaacob is currently a technology consultant at American Express, where he has been working since 2018. He is a graduate of Shizuoka University, Japan, having worked in the Laboratory of Computational and Subjective Image Quality in the Department of Electrical and Electronic Engineering. His research interests include image enhancement and visual perception of texture.

Shun-nan Yang is the director of Vision Performance Institute at the Pacific University College of Optometry. His research specialty is the cortical and subcortical control of eye movements in complex tasks, such as reading and scene viewing. He studies how visual information is processed in the cortex and utilized to initiate cognitive control of eye movement in complex visuomotor sequences. He has published peer-reviewed articles in electronic imaging, vision, optometry, and neurophysiology-related journals.

Yi Zhang received his BS and MS degrees in electrical engineering from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2011, respectively, and his PhD in electrical engineering from Oklahoma State University, Stillwater, Oklahoma, USA, in 2015. From 2016 to 2018, he was a postdoctoral research associate with the Department of Electrical and Electronic Engineering, Shizuoka University, Japan. Currently, he is a faculty member with the School of Electronic and Information Engineering, Xi'an Jiaotong University, China. His research interests include 2D/3D image processing, machine learning, pattern recognition, and computer vision.