Cognition, Technology & Work (2020) 22:585–601
https://doi.org/10.1007/s10111-019-00582-3

ORIGINAL ARTICLE

The effect of labeling on the perceived quality of HDR video transmission

Peter A. Kara[1] · Aron Cserkaszky[2] · Maria G. Martini[1] · Laszlo Bokor[3] · Aniko Simon[4]

Received: 18 February 2019 / Accepted: 27 June 2019 / Published online: 10 July 2019

Abstract
At the time of this paper, high dynamic range (HDR) visualization has already emerged in both the industry and the commercial sector, with HDR displays already present on the consumer market and the technology marching towards the goal of becoming the de facto format of multimedia. HDR is thus often looked at and praised as the next logical step in the evolution of audiovisual entertainment. However, there is no such thing as a single, universal HDR standard, and the competing market projects a future with even more diversity in format specifications, display capabilities and content characteristics. As the competitors attempt to surpass each other and obtain a bigger share of the global market, they inevitably bombard the potential customers and users with brief but effective labels that reflect excellence and superior quality. In this paper, the cognitive effect of such labels is investigated. As home video entertainment is possibly the most numerous instance of the future usage of this visualization technology, video streaming is particularly addressed. Since real-time video transmission services are bound to suffer playback interruptions upon insufficient data rates and uncompensated drops in the available bandwidth, stalling events experienced in conventional multimedia streaming shall apply to HDR video as well. The paper presents four separate experiments, studying how the cognitive bias caused by the labeling effect influences the perception of HDR quality aspects and stalling events, and how the cognitive load varies for stalling detection thresholds between conventional and HDR visualization.

Keywords High dynamic range · Quality of Experience · Cognitive bias · Video stalling

1 Introduction

In the past couple of decades, screens emerged everywhere in our lives. We watch them, we interact with them and, as time progresses, we spend less and less time without them. It is common knowledge that the vast majority of information obtained by a human being is visual and the mobility of screens (i.e., smart phones and tablets) shifts this distribution between our senses even more towards it. Speaking of the mobile devices of the present day, we could say that screens appear in many different shapes and sizes, and the specific shape and the size of a given screen are defined by the scenario in which it is utilized.

We use screens for two primary reasons: for professional and for personal usage. The workplace now has more screens than ever; screens are everywhere in office spaces, we use them to control heavy machinery and to evaluate medical conditions, and the length of this entire paper would not be enough for the sole purpose of listing all the occupations in which screens are used. It is a simple fact which enormously

The work in this paper is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant agreement no. 643072, Network QoE-Net.

✉ Peter A. Kara
p.kara@kingston.ac.uk

Aron Cserkaszky
cserkaszky.aron@itk.ppke.hu

Maria G. Martini
m.martini@kingston.ac.uk

Laszlo Bokor
bokorl@hit.bme.hu

Aniko Simon
aniko.simon@sigmatechnology.se

[1] Kingston University, Kingston upon Thames, London, UK

[2] Pazmany Peter Catholic University, Budapest, Hungary

[3] Budapest University of Technology and Economics, Budapest, Hungary

[4] Sigma Technology, Budapest, Hungary

increases the relevance and importance of human–computer interaction (HCI) research—the more efficient we interact with computers the better. HCI research is of course just as relevant and important for personal usage as well. It involves everything from a laptop to the touch interface of a modern oven. Summa summarum, both professional and personal environments are swiftly accumulating intelligent devices—devices that humans are meant to interact with—and this tendency only seems to accelerate even more every time a novel technology emerges.

The relevance of HCI-related research efforts is particularly boosted by the advances in display technology. There are so many to scientifically address, from the continuous increase in pixel number—i.e., Ultra-High Definition (UHD) displays—to the different forms of 3D apparatus. This paper specifically deals with high dynamic range (HDR) visualization.

Compared to the conventional low dynamic range (LDR) technologies, HDR displays offer visual experience with more realism and immersion. While HDR is rapidly spreading among the instances of 2D visualization, it is also approaching the most novel 3D technologies as well, such as stereoscopic 3D (Yang et al. 2012), virtual reality (Najaf-Zadeh et al. 2017), multi-view displays (Wang et al. 2017), and light field displays (Doronin and Barsi 2018). Every advance we see today in research efforts and display developments bring closer a future where HDR is *the* default screen format.

However, in a way the term “the default format” does not actually exist, as even though HDR displays have just recently appeared on the consumer market, there are already several types and standards. Potential customers, nowadays, meet labels on large displays such as HDR10, HDR10 Plus, and Dolby Vision, and the three-letter acronym of HDR also appears among mobile devices as well. It is expected that the future will bring even more diversity in this context.

But how is this relevant to HCI? The way which we perceive and interact with screens is fundamentally controlled by our cognition. Cognition is a mental process, an intellectual function that either relies on acquired data or on newly created data. Data acquisition is typically executed through our sensory systems—such as seeing or hearing—but new data can also be created via extrapolation and other methods that use existing data. In the framework of HCI, as the process of cognition often determines the interaction with computers (i.e., any intelligent system with input and/or output interfaces that users can engage), cognitive conflict (Dehais et al. 2012) may not only alter the interaction as a whole, but it may also affect the resilience of the system the computer is a part of (Vanderhaegen 2017).

If a person views or interacts with a screen, then the leading cognition is the perception of the screen itself. However, the perception of the screen is not the only cognition in play here. One of the most influential cognition aspects—in the sense of perception-biasing cognition—in this context is expectation, which creates preconceptions regarding the visual experience. Expectations are based on prior experiences and on additional new information. Such information can be any attribute of the screen, or even just the one- or two-word name of the system or the format.

In the research area of Quality of Experience (QoE), this type of cognitive bias is commonly known as the labeling effect. In this interdisciplinary paper, we investigate the effect of labeling on the perceived quality of HDR visualization. More precisely, we study how a single word may affect the perception of HDR video quality. The research simultaneously addresses professional and personal use case scenarios, as the utilization of video transmission is highly spreading in multiple fields of work (e.g., e-health).

As stalling duration is one of the most important quality indicators of on-demand real-time video transmission, the effect of the label on the perceived stalling duration of HDR videos is investigated as well. Furthermore, the research is extended by two works on subjective stalling detection: one that measures the perceptual thresholds of the stalling events used in the research on stalling duration, and one that analyzes the same in the context of conventional LDR visualization, to address the topic of cognitive load and visual attention.

The label in the research was consistently “Premium HDR”, which was compared to simple, regular “HDR”. Such choice for the label was motivated by the common presence in the market of label prefixes like “ultra” (look no further than UHD), “super”, “mega”, and many other words that aim to suggest superior quality and excellence.

The remainder of the paper is structured as follows: Sect. 2 reviews the related work regarding HDR QoE and the labeling effect. Section 3 introduces the apparatus used in the experiments and where the tests took place. Section 4 describes the source videos and their selected stalling event locations. The research on HDR quality aspects, HDR stalling detection, LDR stalling detection, and HDR stalling duration is presented in Sects. 5, 6, 7, and 8, respectively. The paper is concluded in Sect. 9.

2 Related work

The works of Narwaria et al. (2014, 2015a, 2016) address HDR QoE, taking into consideration immersion, the natural feeling of the visualized content, visual attention, and many more aspects, while also discussing subjective measurement methodologies. The authors particularly investigated tone mapping operators (TMOs) and how they affect the perception of HDR content, and also proposed a novel objective video quality metric for HDR (Narwaria et al. 2015b).

Trivially, the major added value of HDR visualization from a QoE perspective originates from the *high* dynamic range itself. However, measuring the dynamic range perceived by test participants is quite far from being a trivial task. The work of Hulusic et al. (2016) introduces a subjective measurement methodology for the perceived dynamic range. The authors carried out a series of subjective tests with 20 test participants, in which HDR images (photographs and video frames) from various sources (e.g., Fairchild's HDR Photographic Survey Fairchild 2007, the Stuttgart HDR Video Database Froehlich et al. 2014, etc.) were assessed on a Full HD (1920 × 1080) SIM2 HDR display, namely the HDR47ES4MB. All still image stimuli were converted to grayscale, as the research solely focused on the perceived dynamic range. The test participants had to evaluate "the overall impression of the difference between the brightest and the darkest part(s) in the image" using a variation of the Subjective Assessment Methodology for Video Quality (SAMVIQ) (Blin 2003). The ratings were collected on a continuous scale (from 0 to 100), which was divided into five labeled, uniform intervals ("*very low*", "*low*", "*medium*", "*high*", and "*very high*"). The findings highlight the importance of content characteristics, such as the relative surface of bright areas and the distance, the separation between dark and bright areas.

Although one of the key features of HDR visualization is the higher level of brightness, having a screen that is too bright might not be preferable by the end user. The work of Bist et al. (2017) proposes a content-based method for brightness control, based on subjective studies of brightness preference. The algorithm operates on a pixel level; the "bright" pixels of the visualized content are taken into consideration during brightness adjustment, which means that the larger the portion of bright areas on the screen, the lower the level of brightness that shall be set. In their experiment, 16 test participants viewed static images on a SIM2 HDR47ES4MB HDR display, the brightness of which they had to re-adjust in case they found the images too bright.

Using physiology in QoE studies is a very well-known approach within the scientific community (Engelke et al. 2017). Depending on the methodology, subjective tests may provide an immense amount of useful information regarding the personal quality preferences and the specific perceptual thresholds of the test participant; however, opinion scores do not report anything about the internal physiological levels of the individual. The work of Al-Juboori et al. (2017) used electroencephalogram (EEG) to analyze the correlation between the perceived quality of HDR images and the different bands of brain activity. Four tone mapping algorithms were applied to five source HDR images and the 20 stimuli were shown to the 28 test participants on an iPhone 6. The results highlight the emotions that were induced by the visualized content, as they correlate with the acquired EEG signals. EEG and peripheral physiological signals were also used by Moon and Lee (2015a, b), who found statistically significant differences in physiological signals between test scenarios of LDR and HDR visualization. EEG was also used by Darcy et al. (2016), and the experiment of Daly et al. (2018) studied pupil behavior during HDR video.

Regarding the labeling effect, there is an abundant literature on this phenomenon of cognitive bias, as it regularly appears in our everyday lives. In such scientific work, information provided to the test participants creates a cognition that may reach a conflicting state with perception and genuine experience, resulting in the need for cognitive dissonance reduction (Festinger 1962). During this process, any cognition—including human experience or its memory—can get altered. Labels include brand, price, and any parameter that may be sufficiently meaningful for the test participant to have a preconception about quality before the actual experience itself. Preconceptions fundamentally depend on the prior experience and related knowledge of the individual, and therefore, the outcome of the labeling effect may deviate greatly and sometimes can be challenging to predict. However, the continuous flow of scientific contributions in the field helps us to understand such cognitive processes, results of which are highly supportive for service and product development, human–computer interfaces, and the integration of novel technologies.

The work of Johansson (1989) addressed how the country of origin (the so-called "Made in" labels) influences consumers, and (Hamzaoui and Merunka 2006) separately investigated the country of design and country of manufacture. Heisey (1990) particularly investigated clothing, taking also into consideration the vendor, the fiber content, and the care procedure. Note that the experiment used identical clothes, and only the label varied. Regarding consumables, (Jacoby et al. 1971) studied the test participants' experience of beer consumption, influenced by the brand of the beer, its composition, and its price. Masson et al. (2008) addressed the influence of the alcoholic content presented on the labels of wine bottles, and (Lick et al. 2017) highlighted the connection between the colors of wine labels and the expectations regarding taste. Verbeke and Viaene (1999) focused on the labeling of beef, and (Burton et al. 1994) generally worked with nutrition reference information.

The presentation of information can play a significant role as well in the overall cognitive bias. The study of Gächter et al. (2009) presented two groups the same information regarding a conference registration fee: pay amount x before a given deadline and pay y after it. However, while one group was told that x was a discount, they introduced y as a penalty towards the other group. When the word "discount" was used, only 67% registered before the deadline, but when a "penalty" was mentioned, this value for the other group was 93%.

Closer to the topic of computer science, Rieh and Belkin (1998, 2000) investigated the effect of domain suffix on the subjective credibility of online information. The research concludes that many scholars were heavily influenced by the suffix of the site, in several cases consciously making decisions purely based on the suffix. Lamm et al. (2010a, b) addressed the performance of search engines and how labeling affects them. The conductors of the subjective tests told half of the participants that it was a professional (and quite expensive) search engine, and told the other half that it was a simple student project. The test participants were further divided by providing half of them with a search engine simulating high system performance, and the other half with low system performance (by intentionally inserting irrelevant results in the system output). Therefore, a given test participant belonged to one of the four possible groups, and only rated one specific search engine with a certain label. The tests were repeated with direct comparisons, where identical or different search engines were to be compared while baring the previously introduced labels.

Sackl et al. (2012a, b) and Kara et al. (2013, 2014) both dealt with the type of Internet connection as the label and addressed different services (browsing and streaming, respectively), and their joint research effort (Kara et al. 2015) investigated the influence of mobile device brand over transmission quality. In a different work of Kara et al. (2017), the label was the resolution of the content (HD or UHD). In this series of subjective tests, beyond the case of stimulus pairs having identical video sequences (either both HD or both UHD) with different labels (one was HD and the other one was UHD), objectively different stimuli were compared as well. Furthermore, half of the test conditions did not contain misleading labels and, thus, reflected the actual resolution of the videos. The analysis of the obtained results indicates that the labels affected the quality ratings of the test participants significantly more than the genuine visual differences.

The previous examples used mock-up methodologies and misinformed test participants with certain labels while presenting test participants typically identical contents, products, and services. The effect of labeling was purely measured via rating scales and feedback forms. Bouchard et al. (2012) also used a similar approach when addressing the sense of presence in virtual reality. Before the subjective tests, one group was told that they were about to be immersed live in a real-time replica of a nearby room, while, for the other group, terms like "live" and "real-time" were completely left out. The aim of the study was to determine whether the term "real-time" enhances the sense of presence or not. This notion was highlighted to the test participants by telling them that the mouse they saw in its cage was also being captured "live" (just like the rest of the room, which was static). Beyond the subjective feedback, the methodology was extended with the involvement of simultaneous functional magnetic resonance imaging (fMRI). Both the subjective results and the fMRI data indicate that the first group was significantly more immersed, yet both groups were presented the exact same synthetic virtual environment.

In this paper, we present two experiments regarding the effect of labeling on the perceived quality of HDR video. Our approach was similar to the previously presented works, as objectively identical contents with different labels were compared. According to the best knowledge of the authors, these studies are the first ones to apply such methodology in the context of HDR visualization. Furthermore, the effect of cognitive bias via labeling on the perceived duration of stalling events has not been addressed in the scientific literature prior to this work, regardless of visualization technology. Accompanying these experiments, we also present two more series of subjective studies on the perceptual thresholds of video stalling detection. The four experiments were also intertwined with each other in the sense that the experimental configuration of one test was based on the results of another. The tests used the same displays and source materials, which are presented in the next sections, followed by the experiments themselves, separately presented in different sections, each with their own research questions, test conditions, and results.

3 Displays and research environments

3.1 HDR research

The subjective tests were performed in an isolated, controlled laboratory environment at the Center of Augmented and Virtual Environments (CAVE) of Kingston University. The ambient luminance was nearly 10 lux and not lower, to avoid visual discomfort (Bist et al. 2017). The test participants viewed the HDR videos on a SIM2 HDR47ES6MB HDR display,[1] with peak brightness over 6000 nits. Behind the display, a D_{65}[2] grey curtain reached from floor to ceiling, serving as background.

The viewing angle was zero degrees (center view) during the entire test, and the viewing distance was a fixed 3 H (1.75 m) according to the recommendation,[3] as Full HD

[1] SIM2 HDR47ES6MB display: http://hdr.sim2.it/hdrproducts/hdr47es6mb.

[2] Rec. P.910: Subjective video quality assessment methods for multimedia applications.

[3] Rec. BT.710: Subjective assessment methods for image quality in high-definition television.

(1920 × 1080) content was displayed on the full screen of the 47-inch Full HD display.

3.2 LDR research

The LDR research was carried out under similar environmental conditions and experimental methodology, located at the Budapest University of Technology and Economics. The only notable difference was the display itself. For these subjective tests, a Panasonic TX-P42S10E was used, which is a 42-inch Full HD plasma television. Similarly, the Full HD content was displayed on the entire screen, but, as the display had a smaller screen compared to the one used in the HDR tests, the 3 H distance in meters (1.57 m) was adjusted accordingly.

4 Source videos

The contents were selected from the Stuttgart HDR Video Database (Froehlich et al. 2014). Table 1 shows the list of the ten chosen contents (see also Fig. 1), their associated IDs, and the starting frames, from which the subsequent 500 frames were cut into 10-bit videos with 24 fps. Source video 2 (“Bistro”) contains one cut and 5 (“Fireplace”) fades from one camera image into another, while the other videos are continuous shots, either with a fixed-position or a panning camera.

The stalling events of the experiments on stalling detection and duration were implemented as frame freezing without visual indicators (e.g., rotating rebuffering icon); the selected frame was shown multiple times (12 times for 500 ms and 24 times for 1000 ms of stalling) before continuing with the next frame. For each content, three stalling event positions were selected, based on their Temporal Information (TI) values, which is a good estimation of the changes between frames. Figures 2 and 3 depict the TI charts of the ten contents defined in Table 1, as well as the positions where frame freezing starts. The first stalling event in every video is denoted as A, the second as B, and the third one as C. Note that, in all three experiments containing impaired videos, a given stimulus always contained exactly one stalling event. One stimulus is identified by the naming convention of either $\{Source\_ID\} + \{Stalling\_event\}$ or $\{Source\_ID\} + \{Stalling\_event\} + \{Stalling\_duration\}$, e.g., in the research on stalling detection, where only one given stalling duration was used, 5C denotes the third stalling event in content 5, and in the research on stalling duration, this is extended with either an S (short duration) or an L (long duration) character, and thus, the identifiers 5CS and 5CL were used.

Table 1 Source video contents used in the HDR experiments

ID	Video content name	Starting frame
1	Beerfest lightshow	102351
2	Bistro	091397
3	Carousel fireworks	097209
4	Cars longshot	092355
5	Fireplace	092341
6	Fishing longshot	060033
7	Poker fullshot	045787
8	Poker travelling slowmotion	033800
9	Showgirl 1	235636
10	Smith welding	248520

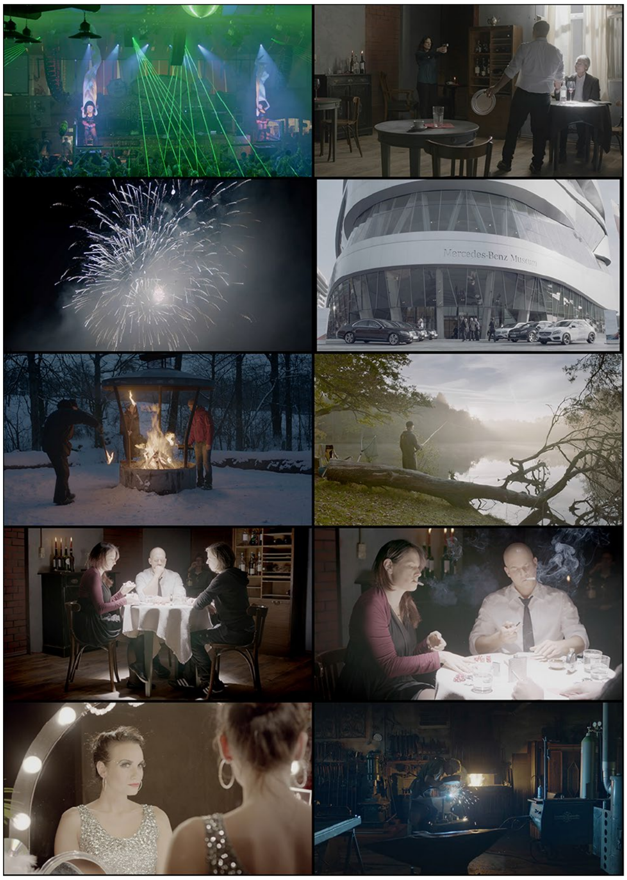

Fig. 1 Source videos used in the subjective tests

The stalling events were particularly positioned on local and global minima and maxima in the TI chart, but also addressed near-identical TI values (even within a content, e.g., 6B and 6C). Some of these events were extreme cases, such as 3C, as shown in Fig. 4; the frozen frame (282) was a 1-frame flash of light. The first and last nearly 2 s of the content were kept clear of stalling.

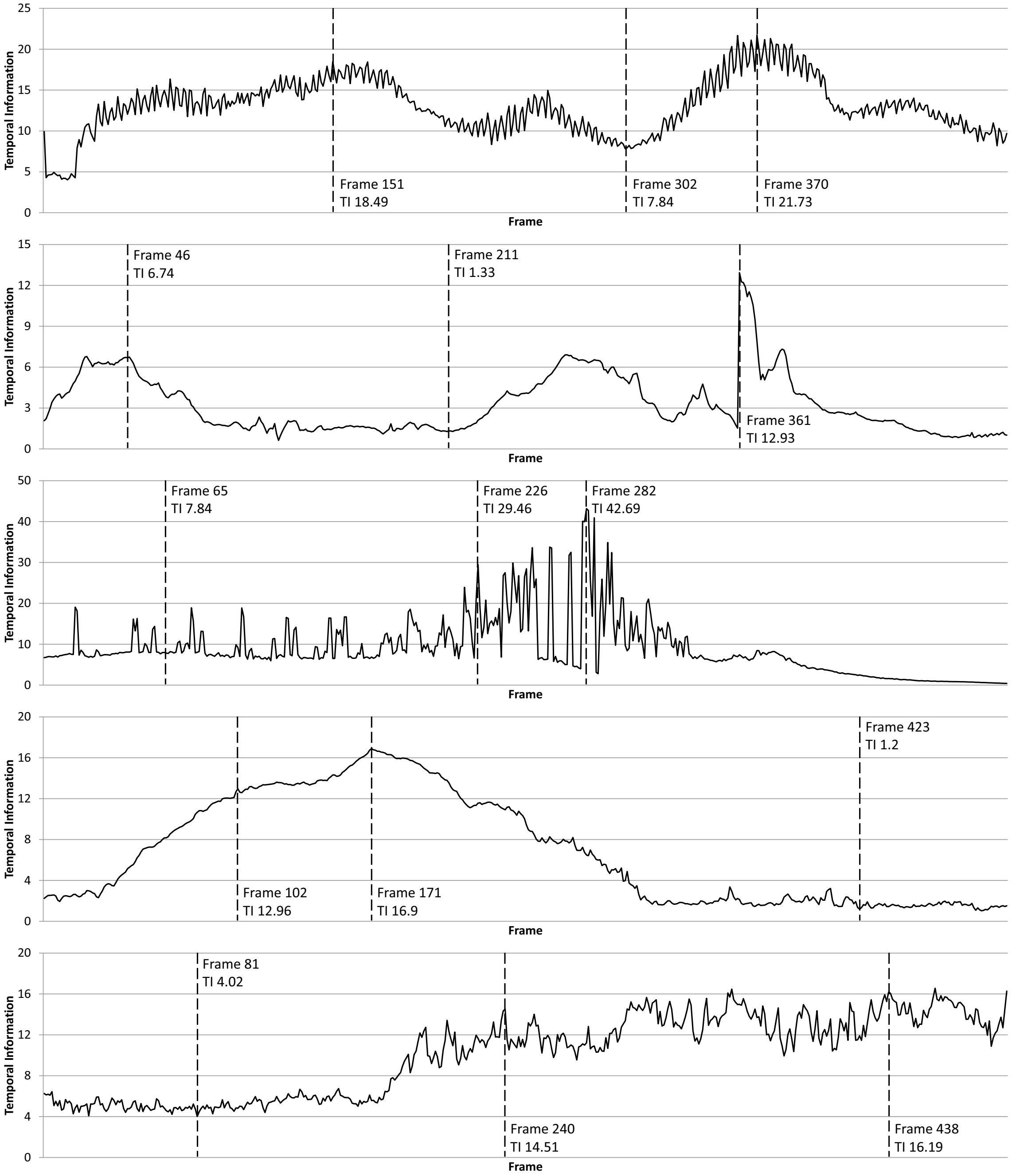

Fig. 2 Temporal information of contents 1–5, presented in a top–down order. The stalling events are denoted with dashed lines

5 Research on HDR quality aspects

5.1 Research question

The aim of the research was to assess the impact of the labeling effect on the selected visual quality aspects.[4]

5.2 Test conditions

The test itself was a pair comparison, which compared video stimuli on a seven-point comparison scale ("*Much worse*", "*Worse*", "*Slightly worse*", "*Same*", "*Slightly better*", "*Better*", and "*Much better*"). To gain a more detailed insight into the cognitive bias created by the labeling effect, instead of comparing the overall QoE, the test participants had to assess four aspects of HDR video quality: *luminance*, *frame rate*, *color*, and *image quality*.

[4] The preliminary findings of this experiment were disseminated in the proceedings of the 2018 International Conference on Multimedia and Expo (ICME) (Kara et al. 2018) prior to this journal paper.

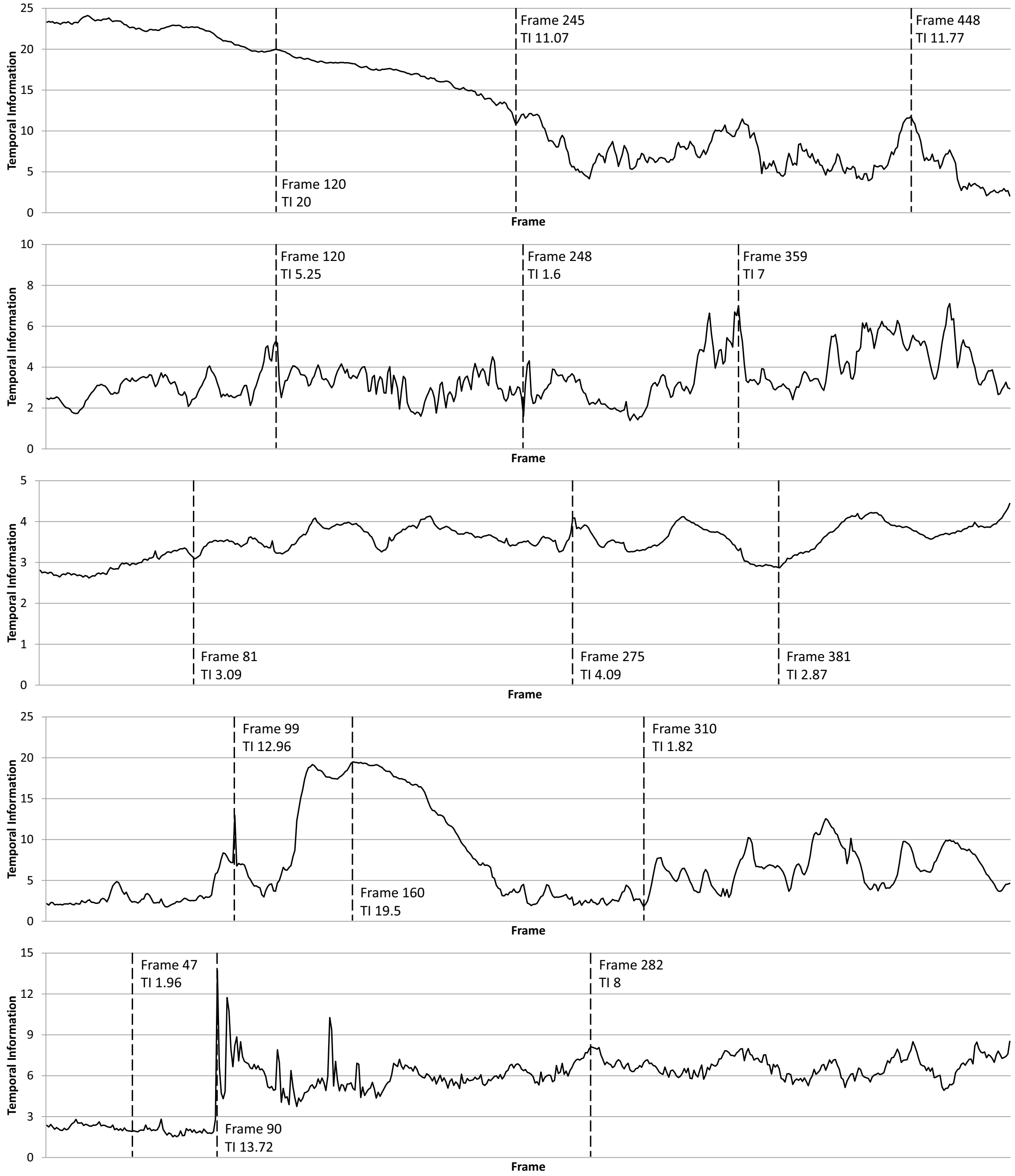

Fig. 3 Temporal information of contents 6–10, presented in a top-down order. The stalling events are denoted with dashed lines

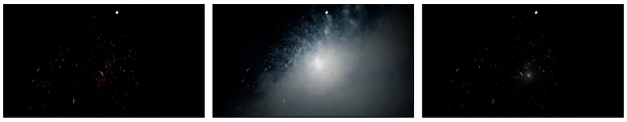

Fig. 4 Frame 281, 282, and 283 of content 3

Before the subjective test, the test participants received training, during which the four aforementioned aspects were interpreted and demonstrated. *Luminance* was described as the perceived difference between the brightest and the darkest portions of the screen; greater difference was to be evaluated better. Although *frame rate* was considerably self-explanatory, it was still explained to every participant to avoid confusion and misunderstandings. *Color* was interpreted as the richness and the depth

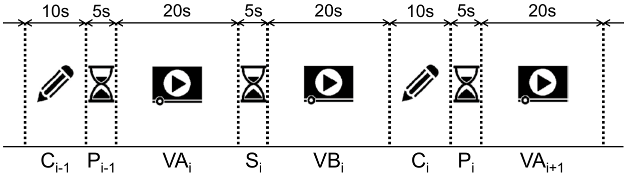

Fig. 5 Temporal structure of the subjective test on quality aspects

of the colors on the screen. Finally, *image quality* was approached from the angle of spatial resolution and classic coding artifacts, independently from the other three aspects.

The double-stimulus method was used, with the stimuli in a pair shown after each other. They were separated by a 5-s blank screen, and comparison was performed directly after each pair, in a time window of 10 s. The stimulus pairs were also separated by a 5-s blank screen.

As detailed in Fig. 5, for a given content *i*—where *i* is a content identifier between 1 and 10, corresponding to the source order randomized for each participant—the first instance of the content (VA_i) is played, followed by the stimulus separation (S_i), and then, the identical second instance (VB_i) is shown. After this, VB_i is compared to VA_i in the comparison period (C_i), and finally, the separation screen between the pairs is displayed (P_i). As this given structure is repeated over the duration of the subjective test, if *i* is at least 2 but at most 9 (i.e., neither the first not the last pair), then VA_i occurs directly after the comparison period and the separation screen of the prior content $i - 1$ (C_{i-1} and P_{i-1}, respectively), and P_i is followed by the first instance of the subsequent content $i + 1$ (VA_{i+1}).

The order of stimuli in the subjective quality assessment varied among test participants. For half of the participants, the "Premium" video was always the first one in the pair (*VA*), and for the other half, it was the second one (*VB*). Again, this means for each and every test participant, that the assignment of the label was consistent and did not change during the test. As the labeling effect can influence both perception and the memory of perception, this given division between the test participants was included to investigate the role of label order.

5.3 Results

A total of 40 individuals participated in the tests (30 males and 10 females). The age range was from 20 to 56, and the average age was 30. Ten participants had prior HDR video experience, and the rest had never seen any HDR video before the experiment.

The obtained subjective scores are represented by their numerical counterparts, ranging from −3 to +3. During the subjective tests, the test participants were presented a combination of the available qualitative tags for stimulus comparison—defined in the previous section—and these values, emphasizing a uniform distance between the values of the scale. In this analysis, positive values favor the "Premium HDR" stimulus, while negative values indicate that it was deemed to be worse in the given aspects.

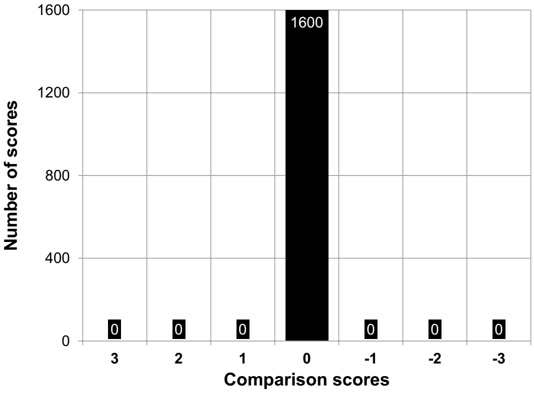

Fig. 6 Ideal distribution of scores of the subjective test on quality aspects

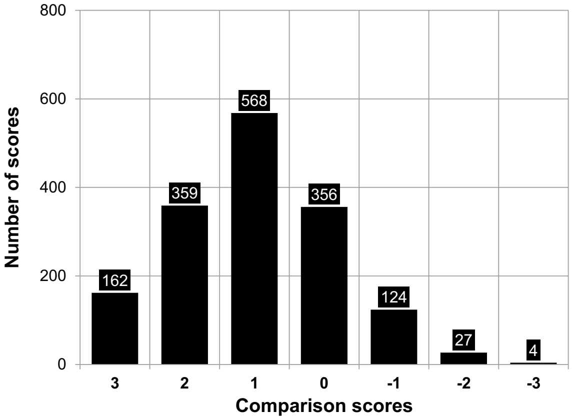

Fig. 7 Scoring distribution of the subjective test on quality aspects

Each of the 40 test participants compared 4 quality aspects of 10 stimulus pairs, and thus, 1600 subjective scores were collected in the experiment. In an ideal scenario without the presence of cognitive bias through the labeling effect (and of course without any other type of subjective bias), all these 1600 scores would have reported the given aspects to be the "*Same*" (see Fig. 6). However, according to the scoring distribution, only 356 (22.25%) of them were zero, and 1244 (77.75%) assessed a certain level of either positive (1089 scores) or negative (155 scores) difference (see Fig. 7).

The most frequent quality comparison score was "*Slightly better*", followed by "*Better*", "*Same*", "*Much better*",

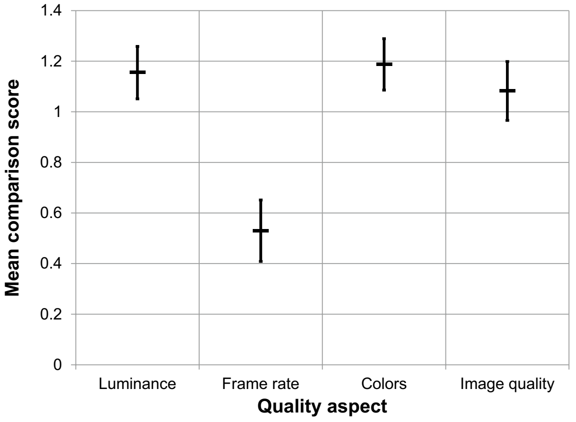

Fig. 8 Mean comparison scores of the subjective test on quality aspects

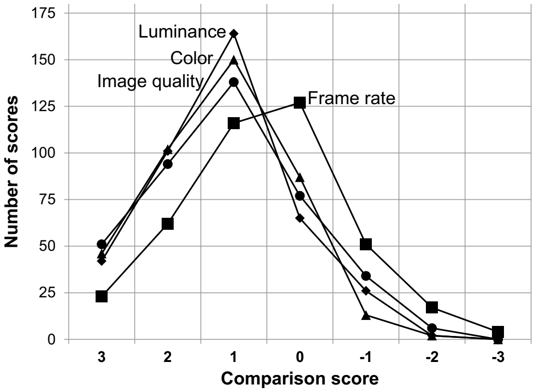

Fig. 9 Scoring distribution of the quality aspects

"*Slightly worse*", "*Worse*", and "*Much worse*". This order took all of the investigated aspects into consideration. If we separate them, we can observe rather similar mean values for *luminance*, *color*, and *image quality* (see Fig. 8). In fact, the aforementioned order in score frequency applied to all three of them (see Fig. 9) and there was no statistically significant difference between them.

However, *frame rate* was assessed differently. The mean score was significantly lower compared to the other aspects, as the number of positive scores was the lowest, while it received the most zero and negative scores. Moreover, the number of negative scores *frame rate* received was near to the number of negative scores received by the other three aspects together.

It needs to be noted that more than half (201 out of 400) of the scores for *frame rate* were, indeed, positive, meaning that the test participants providing those scores experienced an improvement in this aspect for the stimuli with "Premium HDR" quality. Yet, there were many who either did not perceive a change in *frame rate* or experienced degradation.

Although the experimental setup did not define any feedback beyond the comparison scores, some test participants provided us valuable insights to their visual experience. One of the test participants, who works in the movie industry, claimed that

> "The first version (Premium HDR) is always more pushed to the limits; it's actually more magical, but less controlled. The second one (HDR) feels more controlled, less magic. Personally I would go for a middle path. The frame rate doesn't seem to improve significantly."

There were also test participants who consistently experienced frame rate drops in the "Premium HDR" videos, while perceiving improvements in the other aspects. Their comparison patterns can be summarized by the following feedback:

> "It is such a pity that these incredible visuals come at the expense of frame rate. Yet to be fair, it is most certainly worth it."

The cognition originated from the concept of compensation, the idea of balance; if certain aspects become better, then their improvements negatively affect the performance of others. One could suggest that such bias might be limited to test participants with educational backgrounds of engineering or computer science, but these patterns appeared randomly within the observer population. The impressive visuals of HDR compared to regular LDR TV experience are easier to connect with a "premium" quality when it comes to *luminance*, *color*, and even *image quality*, compared to a *frame rate* of 24 fps, when 60 fps is spreading in the everyday use case scenarios. In addition, from the three highlighted aspects, *image quality* received the least positive and the most negative scores, even though it was not statistically different from the other two. Repeating the same experiment in UHD resolution is expected to boost this aspect in the positive direction.

Regarding the effect of the label order, no significant difference was found between the ratings of the two groups and the general findings applied to this scoring separation as well. When statistically analyzing the data for each source content, the one and only case for which a significant difference was found was the *image quality* of source video sequence 1. When the "Premium HDR" was the first stimulus, the mean was 1.4, but when it was the second one, it was only 0.7. For this comparison, the p value of the ANOVA test was 0.012. For the other 39 cases, it was above 0.05, and for 27 comparisons, it was above 0.5, even reaching 1 (e.g., *image quality* of content 3 or *color* of content 1). Therefore, based on these results, the influence of the order of labeling was not investigated in further experiments.

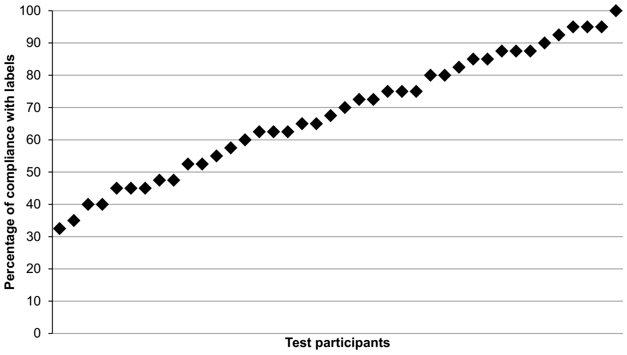

Fig. 10 Percentage of compliance with labels in the subjective test on quality aspects

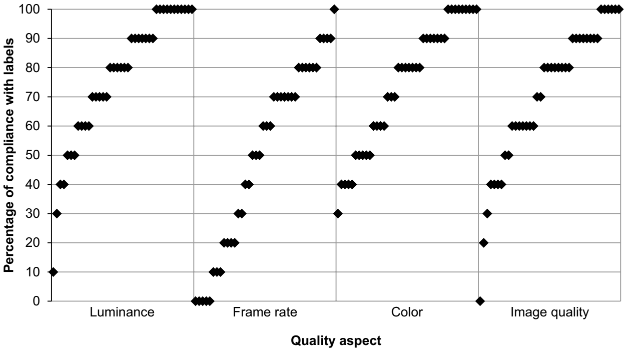

Fig. 11 Percentage of compliance with labels per quality aspect

Finally, the compliance with labels was measured. In the context of this experiment, the decision of the test participants was considered compliant to the labels, if the "Premium HDR" stimulus was preferred (positive ratings). The overall compliance and the per-aspect compliance are shown in Figs. 10 and 11, respectively. In both figures, each marker represents the rate of compliance for a test participant. The results indicate a rather even distribution between 30 and 100% of compliance rate, with an average rate of 68.06%. In this analysis, 100% means that the test participant preferred the "Premium HDR" stimulus for each and every source sequence and quality aspect. This applied only to a single test participant. When separated by quality aspect, we can see that a 100% of compliance was achieved by 11, 9, 6, and 1 test participants for *luminance*, *color*, *image quality*, and *frame rate*, respectively. The average rates of compliance for this order of quality aspects were 76.75%, 74.5%, 70.75%, and 50.25%. Note that, in case of *frame rate*, 5 out of 40 test participants achieved a rate of 0%, which means that they either did not distinguish the stimuli of assessed the *frame rate* of the regular "HDR" stimulus to be better. All things considered, the high average rates for the three other quality aspects further reinforce the findings on the influence of the labeling effect.

6 Research on HDR stalling detection

6.1 Research question

The aim of the research was to assess the perceptual sensitivity towards a stalling event with a given duration on an HDR display.

6.2 Test conditions

The subjective test was performed using a double-stimulus methodology for a pair comparison with a five-point Degradation Category Rating (DCR) scale ("*Imperceptible*", "*Perceptible but not annoying*", "*Slightly annoying*", "*Annoying*", and "*Very annoying*"). For every test condition, the test participants compared an impaired stimulus (containing a single stalling event) to the reference video. They had to assess whether the playback interruption was observable or not and, if it was, then how annoying it was (as defined by the scale).

Instead of focusing on perceptual thresholds based on stalling duration—which has already been extensively investigated in the past—the primary focus was on the content itself through TI. Thus, one single stalling duration was used for every test condition, and the stimuli only varied in content and the positioning of the event. The duration of 500 ms was chosen, which is, according to the literature, a clearly perceivable duration (van Kester et al. 2011; Usman et al. 2015; Staelens et al. 2010; Yu et al. 2015; Kara et al. 2016a, b). The test participants were not aware that the stalling duration was the same in every stimulus.

After the training phase, the stimuli pairs were shown in random order, and were separated by 5-s blank screens. The rating task was performed directly after each impaired video stimulus. As there were 3 stalling events for 10 source videos, this means that 30 stimuli were to be assessed, each with the duration of 21.3 s (512 frames).

The subjective test was followed by a post-experiment questionnaire. These questions addressed the memory bias, as test participants had to recall attributes of the stimuli which they did not focus on. They were asked about the perceived variation about the aspects of *luminance*, *frame rate*, *color*, and *image quality*. Prior to the experiment, they were not informed about the questions of the post-experiment questionnaire, as these aspects would have diverted attention away from the stalling events. For each aspect, the test participants were asked whether there was a variation at all, and if there was, the number of affected contents was to be specified. The possible options were "No", "Not sure", "1–3 contents", "4–6 contents", and "7–10 contents".

As the main part of this research focused exclusively on perceptual thresholds, the term "Premium HDR" was not

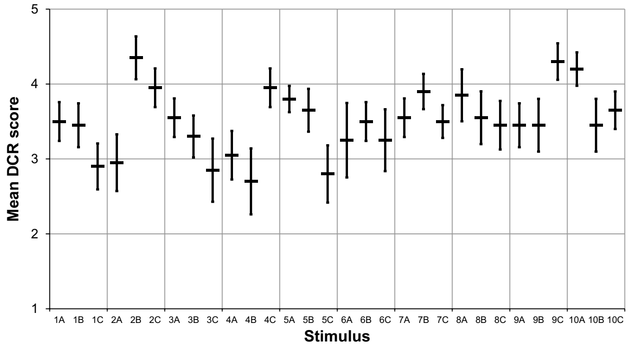

Fig. 12 Mean DCR scores of the subjective tests on HDR stalling detection

used during the test. The same applies to the identical test with LDR visualization.

6.3 Results

A total of 20 individuals participated in the tests (15 males and 5 females). The age range was from 21 to 37, and the average age was 28. Three participants had prior HDR video experience, and the rest had never seen any HDR video before the experiment.

The mean scores are shown in Fig. 12. Although each and every stalling event had the exact same duration (500 ms), the impact on perception varied significantly.

The greatest difference can be observed in case of content 2, between 2A (mean score 2.95) and 2B (mean score 4.35). Again, the stalling duration was identical; however, while 2A was a fast-paced walking motion from the right to the left, across the entire scene, 2B was limited to subtle hand motions. As for 2C, its TI value was nearly twice the value of 2A, yet it received particularly high scores. 2C was at a sudden scene change within the content and, thus, the spike in the TI chart. Stalling was not only well tolerated at this frame, but also eluded the perception of three test participants.

Such cases, when test participants failed to perceive the 500 ms stalling event in the stimuli and provided "*Imperceptible*" as the assessment score, are summarized in Fig. 13, displayed together with the corresponding TI values. According to this analysis, 2B was, indeed, the least noticed, followed by 9C and 10A. These frames are shown in Fig. 14.

Table 2 shows the results of the post-experiments questionnaire. The first things that really stand out from the data are that not a single test participant stated that there was no variation in *frame rate*, and that the number of unsure test participants was by far the lowest as well. In fact, nearly half, 9 out of 20 test participants stated that at 4, 5, or 6 contents contained *frame rate* variations. *Image quality* was clearly the least affected by the memory bias, followed by *color* and *luminance*.

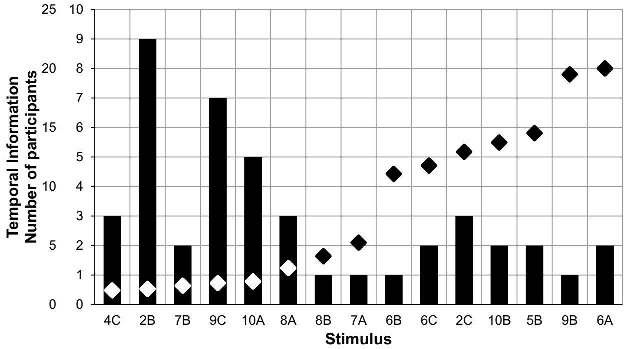

Fig. 13 Number of test participants who assessed the given stimuli with "Imperceptible" ratings (bars) and the TI of the corresponding stimulus (markers)

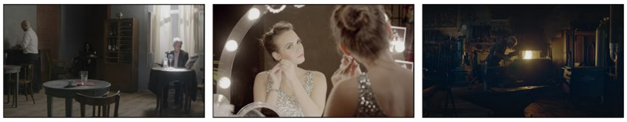

Fig. 14 Frame freezing at 2B, 9C, and 10A

Table 2 Results of the post-experiment questionnaire

	No	Not sure	1–3 Contents	4–6 Contents	7–10 Contents
Luminance	2	6	8	3	1
Frame rate	0	2	7	9	2
Color	4	8	3	2	3
Image quality	6	8	5	1	0

7 Research on LDR stalling detection

7.1 Research question

The aim of the research was to assess the perceptual sensitivity towards a stalling event with a given duration on an LDR display, and, thus, serve as a comparison to the previously introduced experiment.

7.2 Test conditions

The test conditions were identical to the parameters of the research on HDR stalling detection. The differences in the experimental setup compared to the HDR counterpart were

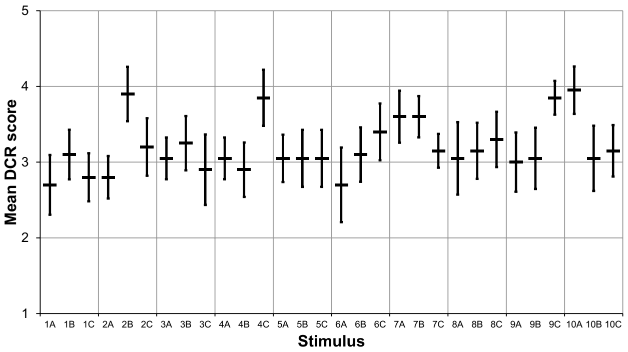

Fig. 15 Mean DCR scores of the subjective tests on LDR stalling detection

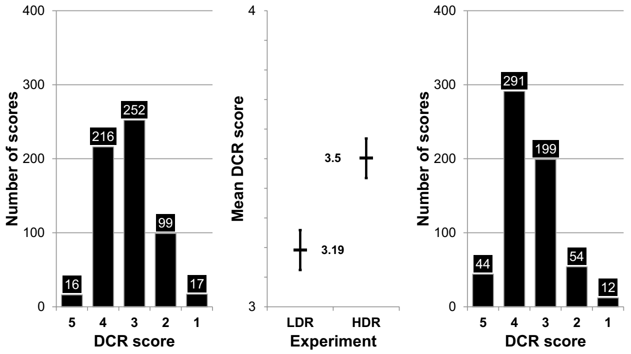

Fig. 16 Scoring distribution of the LDR (left) and the HDR (right) experiment on stalling detection, and their mean scores (middle)

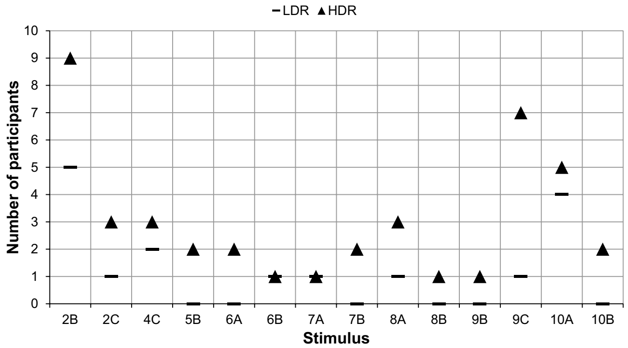

Fig. 17 Number of test participants in the LDR and HDR tests who assessed the given stimuli with "Imperceptible" ratings

only the display itself (as introduced earlier in the paper) and, of course, the bit depth of the stimuli.

7.3 Results

A total of 20 individuals participated in the tests (18 males and 2 females). The age range was from 21 to 60, and the average age was 29.7.

The mean scores are shown in Fig. 15. At first glance, the figure indicates that the obtained scores of several test stimuli were lower than what was achieved for HDR stalling detection, and variations were smaller as well. To be precise, while the average of all HDR scores was 3.5, the corresponding value for LDR was 3.19. This suggests that the stalling events in the HDR experiment were more difficult to perceptually detect and/or they were more tolerable, compared to the LDR experiment. However, in order to draw any conclusion, a direct comparison with statistical analysis is required.

7.4 Comparison of HDR and LDR stalling detection

Figure 16 compares the scoring distributions and the aforementioned means of the two experiments. The latter indicates a significant difference, as the 0.95 CIs do not overlap. This difference is well reflected in the scoring distributions. Since both subjective studies addressed stalling detection, the most important DCR score in this analysis is 5 ("*Imperceptible*"). While the HDR experiment produced 44 of this score, this was only 16 in case of LDR.

Does this mean that compared to conventional LDR visualization, HDR stalling events were more difficult to detect in general? Not necessarily. To gain more insight, let us compare the distribution of these scores particularly. Figure 17 shows the number of test participants using the "*Imperceptible*" rating option for the given test stimuli, separately for LDR and HDR. The results show that every HDR stalling event received as least as many "*Imperceptible*" ratings as LDR did. The greatest differences were measured for 2B and 9C, which were the two least detectable stalling events in the HDR study (see Fig. 13). These findings indicate that difficult-to-perceive stalling events (with minimal amounts of variation between adjacent frames) may go unnoticed during HDR visualization, but the same is less likely to happen in case of LDR.

It is important to note that the findings presented so far do not mean that each and every test stimulus differed significantly. The statistical analysis of the conditions are presented in Table 3. We can see that for 9 out of 30 cases, the difference was statistically significant. In all of these cases, HDR visualization achieved significantly higher scores, and thus, these stalling events were more difficult to detect and/or easier to tolerate. The differences between them on the scale from 1 to 5 were at least 0.35, but, for 8A, it was 0.8. Some of these frame repetitions were rather subtle—like the ones presented in Figs. 13 and 17—while some were quite obvious.

The results of the comparison do not correlate with TI due to the aforementioned diversity, but they are most definitely connected to the so-called "visual awe". Let us take 1A, 5B, and 10C (see Fig. 18) as counter-examples for the

Fig. 18 Frame freezing at 1A, 5B, and 10C

Table 3 Statistical analysis of the conditions (c) of the LDR and HDR stalling detection; each line compares the results of a given test condition for the two experiments

c	*p*	*S*
1A	0.002	Yes
1B	0.123	No
1C	0.657	No
2A	0.538	No
2B	0.059	No
2C	0.002	Yes
3A	0.013	Yes
3B	0.828	No
3C	0.876	No
4A	1.000	No
4B	0.495	No
4C	0.661	No
5A	0.000	Yes
5B	0.016	Yes
5C	0.364	No
6A	0.130	No
6B	0.080	No
6C	0.600	No
7A	0.818	No
7B	0.108	No
7C	0.033	Yes
8A	0.010	Yes
8B	0.131	No
8C	0.547	No
9A	0.075	No
9B	0.148	No
9C	0.011	Yes
10A	0.203	No
10B	0.162	No
10C	0.024	Yes

The *p* value of ANOVA is given (*p*), along with significance (*S*)

idea that hard-to-detect, low-TI stalling events differ more between LDR and HDR visualization technologies. All of these stalling events had high TI values, as shown on Figs. 2 and 3. 1A captured a vertical camera panning during a highly dynamic scene, 5B was a closeup on the lit bonfire with added movement on the right, and 10C also captured camera movement, during the visually intense moment of welding. Therefore, these stalling events were difficult to miss (yet for 5B, two test participants actually managed to, during the HDR test, as shown in Fig. 13), but they were all visually impressive. To be more precise, they were visually impressive when shown as HDR contents on an HDR display.

What was also common in them is that the stalling event itself was not *too* irritating. Let us now examine 3C, with its one-frame flash of light (see Fig. 4). The mean scores for the LDR and the HDR tests were 2.9 and 2.85, respectively, not a single test participant deemed it "*Imperceptible*", only 6–7 found it not to be annoying, and the worst score "*Very annoying*" appeared twice in both experiments. Similar assessments were applied to 3B as well, which also repeated the selected frame amidst sudden flashes, and the achieved means were 3.25 and 3.3. The reason why source video 3 ("Carousel Fireworks") is a good example for the very similar ratings in both experiments, is that it had the greatest contrast due to the pitch-black night sky and the exceptionally bright fireworks. Yet, the test participants were similarly annoyed, regardless of visualization. However, 3A—which was before the bright flashes and, therefore, the visual awe was not disturbed by a highly annoying stalling position—was rated differently for LDR and HDR (means of 3.05 and 3.55, respectively) and, in fact, the difference was statistically significant.

8 Research on HDR stalling duration

8.1 Research question

The aim of the research was to assess the impact of the labeling effect on the perceived duration of stalling events.

8.2 Test conditions

For the indication of difference in perceived stalling duration, a seven-point scale was used ("*Much shorter*", "*Shorter*", "*Slightly shorter*", "*Same*", "*Slightly longer*", "*Longer*", and "*Much longer*"). Based on the results of HDR stalling detection, for each source video, two stalling events were selected: the easiest and hardest to detect and tolerate. Table 4 shows these selected stalling events. Each stalling was included twice, once with a duration of 500 ms and once with 1000 ms. Therefore, each source video was assessed four times, and thus, 40 comparisons were made. Labeling was present in the experiment, in a similar manner as in the research on HDR quality aspects; the utilized mock-up methodology here was the same as before.

8.3 Results

A total of 36 individuals participated in the tests (22 males and 14 females). The age range was from 20 to 42, and the average age was 26. Eight participants had prior HDR video

Table 4 Selected stalling events for the research on HDR stalling duration

	1	2	3	4	5	6	7	8	9	10
A	X	X	X	–	X	–	–	X	X	X
B	–	X	–	X	–	X	X	–	–	X
C	X	–	X	X	X	X	X	X	X	–

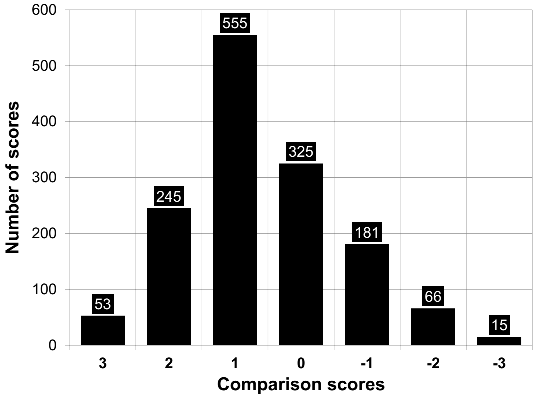

Fig. 19 Scoring distribution of the subjective tests on stalling duration

experience, and the rest had never seen any HDR video before the experiment.

The obtained subjective scores are represented by their numerical counterparts, ranging from −3 to +3. During the subjective tests, the test participants were presented a combination of the available qualitative tags for stimulus comparison—defined in the previous subsection—and these values, emphasizing a uniform distance between the values of the scale. In this analysis, positive values indicate longer perceived stalling durations for the “Premium HDR” stimulus, while negative values indicate that it was perceived as the shorter one.

With 36 test participants and 40 comparisons, a total of 1440 scores were collected. Figure 19 shows the distribution of these scores. It is apparent that the labeling effect had a significant impact on the perception of stalling duration. Only in 22.6% of the ratings indicated no perceived difference between the identical video stimuli, which is very similar to the scoring distribution of the experiment on quality aspects (22.25%, see Fig. 7).

The obtained ratings are decisively positive (59.2%), which means that the stimuli labeled as “Premium HDR” were generally perceived to have longer stalling events. The most common score by far was +1 (38.5%), indicating slightly longer stalling event for “Premium HDR” stimuli. Negative scores are present as well in the analysis (18.2%), but the number of −2 and −3 is particularly low (5.6% combined), while the same cannot be said for the corresponding positive scores (20.7% combined).

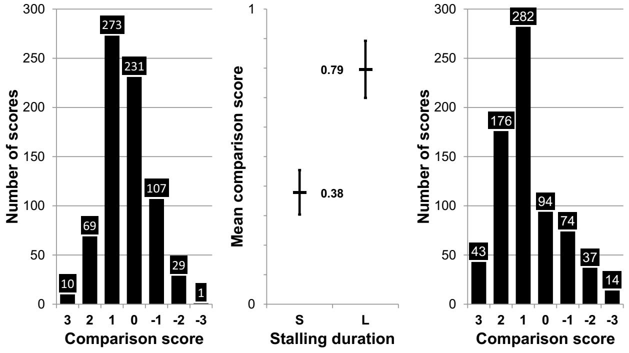

Fig. 20 Scoring distribution of short (left) and long (right) stalling events, and their mean comparison scores (middle)

In this experiment, two different stalling durations were used. Figure 20 shows their separate scoring distributions and their mean comparison scores. The results clearly indicate that the bias in perception was significantly stronger for the video stimuli with longer stalling durations. While 32.1% of the scores of the stimuli with short stallings report the lack of difference, this is only 13% for long stallings.

Figure 21 shows the number of 0 (“Same”) scores for each test stimulus, and the mean comparison scores with 0.95 CI. The highest numbers of 0 scores were achieved by 9CS, 2BS, and 10AS, which were the stimuli with the least detectable/annoying stalling events (see Figs. 12, 13, 14). The findings extracted from Fig. 20 apply here as well, since the upper half of the descending order of 0 scores is dominated by short stalling events. In accordance with the distribution of Fig. 19, mean scores rise with as 0 scores get lower, however, statistical differences are difficult to find, due to the large scoring deviations. Note that it is likely in such experiment that while a specific test participant rates a given stimulus with +3, a different participant may rate it as −3. Standard deviation at the upper end of this order (highest numbers of 0 scores) is only 0.7–0.8, while at the other end, it is 1.6–1.7.

Finally, the compliance with the labels is addressed. As “Premium HDR” is basically a positive label, the compliance rate in this context is based on the ratings that indicate a shorter perceived stalling duration for the “Premium HDR” stimulus. The overall percentage of compliance and the compliance rate separately investigated for the different stalling duration are shown in Figs. 22 and 23, respectively. Similarly to Figs. 10 and 11, each marker represents the rate

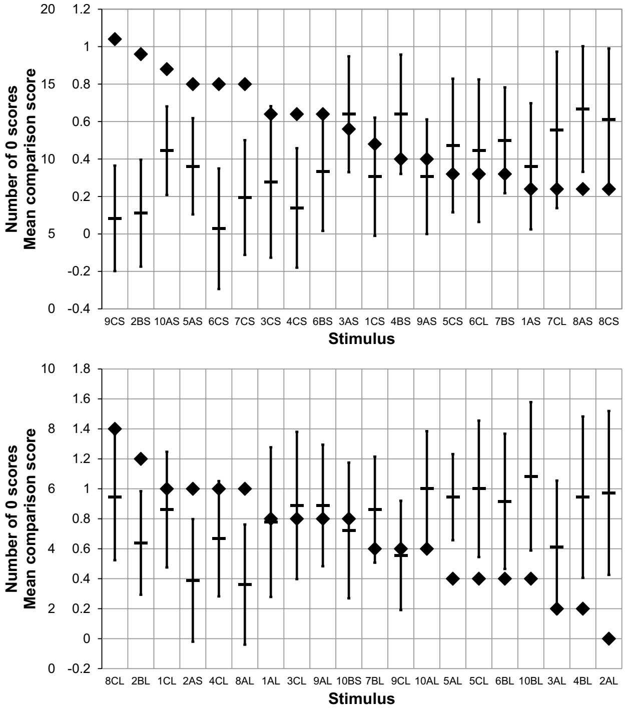

Fig. 21 Number of 0 scores (markers) and mean comparison scores (intervals) of the research on stalling duration

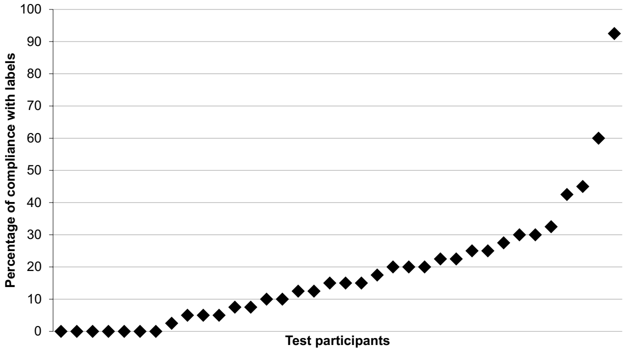

Fig. 22 Percentage of compliance with labels in the subjective test on stalling duration

of compliance for a test participant. The average compliance rate was 18.19%, with 7 out of 36 test participants who never found the stalling event of the "Premium HDR" stimulus to be shorter than the other one. 31 test participants had a compliance rate of 30% or less, and only two overall rates were above 50%. In comparison, 31 out of 40 test participants had the corresponding value above 50% in the experiment on quality aspects, and not a single individual had an overall rate below 30%. Regarding the separation based on stalling duration, the average rates for the short and the long stalling events were 19.03% and 17.36%, respectively, and one test participant reached a 100% rate for the stimuli with the long stalling events. The low compliance rates in general indicate that the vast majority of test participants did not believe that a format with superior visualization quality should have shorter stalling durations. In fact, the common concept (or rather common preconception) was actually the opposite, as shown by the results.

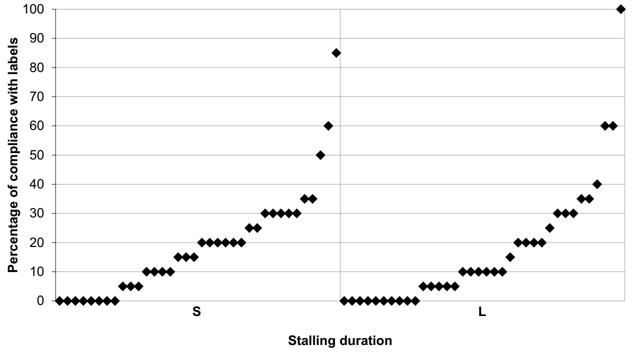

Fig. 23 Percentage of compliance with labels for short (S) and long (L) stalling events

9 Conclusion

The paper introduced four experiments on HDR video QoE. The first one addressed different quality aspects, and investigated their cognitive distortions caused by the labeling effect. The obtained subjective scores indicate that, for aspects like luminance, color and image quality, the positive label "Premium HDR" resulted in a positive bias, but, for frame rate—which was more difficult to directly connect to HDR visualization—the rating patterns were not obvious. It was found that several test participants approached frame rate as an aspect which generally suffers degradations due to a trade-off between visuals and frame rate.

The second and the third experiments focused on stalling event detection and tolerance for HDR and LDR visualization, respectively. The comparison of the results showed that even 500 ms stalling events may go unnoticed due to the presence of the so-called "wow effect" and "visual awe" that comes with HDR visualization. The studies indicate statistically significant differences between the evaluation of the LDR and HDR sequences.

The fourth experiment investigated the perceived duration of stalling events when the test participant is influenced by the label "Premium HDR". The findings clearly suggest the presence of the prior idea; the preconception, which—similarly to the first experiment—builds on the trade-off between visuals and other quality aspects that do not contribute to the appearance of HDR visualization. The results indicate that the stalling events in "Premium HDR" videos were perceived to be longer. This applied to stalling events with 500 ms and 1000 ms duration as well, but the latter suffered significantly more cognitive bias.

As for future work, the experiments can be extended in numerous potential research directions. First of all, the second and the third study can be repeated with several other stalling event durations and patterns (i.e., varying stalling frequencies within a stimulus with different durations), particularly targeting short events around the level of just noticeable difference (JND) and longer durations beyond 1000 ms. Regarding quality aspects, the assessment of frame rate and stalling event duration could be simultaneously integrated into an experiment with the presence of the labeling effect. Furthermore, the fading of cognitive bias over time could also be investigated, with various test methodologies and significantly longer video sequences.

Acknowledgements The authors would like to thank all the test participants and everyone who made the presented research possible.

References

Al-Juboori S, Mkwawa I-H, Sun L, Ifeachor E (2017) Investigation of relationships between changes in EEG features and subjective quality of HDR images. In: International conference on multimedia and expo (ICME). IEEE, pp 91–96

Bist C, Cozot R, Madec G, Ducloux X (2017) QoE-based brightness control for HDR displays. In: Ninth international conference on quality of multimedia experience (QoMEX). IEEE, pp 1–6

Blin J-L (2003) SAMVIQ-Subjective assessment methodology for video quality. Rapp Tech BPN 56:24

Bouchard S, Dumoulin S, Talbot J, Ledoux A-A, Phillips J, Monthuy-Blanc J, Labonté-Chartrand G, Robillard G, Cantamesse M, Renaud P (2012) Manipulating subjective realism and its impact on presence: preliminary results on feasibility and neuroanatomical correlates. Interact Comput 24(4):227–236

Burton S, Biswas A, Netemeyer R (1994) Effects of alternative nutrition label formats and nutrition reference information on consumer perceptions, comprehension, and product evaluations. J Public Policy Mark 13:36–47

Daly S, Gitterman E, Mulliken G (2018) Pupillometry of HDR video viewing. Electron Imaging 2018(14):1–8

Darcy D, Gitterman E, Brandmeyer A, Daly S, Crum P (2016) Physiological capture of augmented viewing states: objective measures of high-dynamic-range and wide-color-gamut viewing experiences. Electron Imaging 2016(16):1–9

Dehais F, Causse M, Vachon F, Tremblay S (2012) Cognitive conflict in human–automation interactions: a psychophysiological study. Appl Ergon 43(3):588–595

Doronin O, Barsi A (2018) Estimation of global luminance for HoloVizio 3D display. In: International conference on 3D immersion (IC3D). IEEE, pp 1–8

Engelke U, Darcy DP, Mulliken GH, Bosse S, Martini MG, Arndt S, Antons J-N, Chan KY, Ramzan N, Brunnström K (2017) Psychophysiology-based QoE assessment: a survey. IEEE J Sel Top Signal Process 11(1):6–21

Fairchild MD (2007) The HDR photographic survey. In: Color and imaging conference, vol 2007, no 1. Society for imaging science and technology, pp 233–238

Festinger L (1962) A theory of cognitive dissonance, vol 2. Stanford University Press, Stanford

Froehlich J, Grandinetti S, Eberhardt B, Walter S, Schilling A, Brendel H (2014) Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In: Digital photography X. SPIE, vol 9023, pp 1–10

Gächter S, Orzen H, Renner E, Starmer C (2009) Are experimental economists prone to framing effects? A natural field experiment. J Econ Behav Organ 70(3):443–446

Hamzaoui L, Merunka D (2006) The impact of country of design and country of manufacture on consumer perceptions of bi-national products' quality: an empirical model based on the concept of fit. J Consum Mark 23(3):145–155

Heisey FL (1990) Perceived quality and predicted price: use of the minimum information environment in evaluating apparel. Cloth Text Res J 8(4):22–28

Hulusic V, Valenzise G, Provenzi E, Debattista K, Dufaux F (2016) Perceived dynamic range of HDR images. In: Eighth international conference on quality of multimedia experience (QoMEX). IEEE, pp 1–6

Jacoby J, Olson JC, Haddock RA (1971) Price, brand name, and product composition characteristics as determinants of perceived quality. J Appl Psychol 55(6):570–579

Johansson JK (1989) Determinants and effects of the use of made in labels. Int Mark Rev 6(1):47–58

Kara PA, Bokor L, Imre S (2013) Distortions in QoE assessment of 3D multimedia services on multi-access mobile devices. In: IEEE 9th international conference on wireless and mobile computing, networking and communications (WiMob), pp 311–318

Kara PA, Bokor L, Imre S (2014) Seeing is believing and vice versa: investigation of the altered perception during subjective assessment of streaming multimedia. In: Tenth international conference on signal-image technology and internet-based systems (SITIS). IEEE, pp 539–545

Kara PA, Bokor L, Sackl A, Mourão M (2015) What your phone makes you see: investigation of the effect of end-user devices on the assessment of perceived multimedia quality. In: Seventh international workshop on quality of multimedia experience (QoMEX). IEEE, pp 1–6

Kara PA, Cserkaszky A, Martini MG (2018) Premium HDR: the impact of a single word on the quality of experience of HDR video. In: International conference on multimedia and expo (ICME), emerging multimedia systems and applications (EMSA). IEEE, pp 1–6

Kara PA, Robitza W, Martini MG, Hewage CT, Felisberti FM (2016a) Getting used to or growing annoyed: how perception thresholds and acceptance of frame freezing vary over time in 3D video streaming. In: IEEE international conference on multimedia and expo workshops (ICMEW), pp 1–6

Kara PA, Martini MG, Hewage CT, Felisberti FM (2016b) Times change, stalling stays: subjective quality assessment over time of stalling in autostereoscopic 3D video services. In: 12th International conference on signal-image technology and internet-based systems (SITIS). IEEE, pp 787–792

Kara PA, Robitza W, Raake A, Martini MG (2017) The label knows better: the impact of labeling effects on perceived quality of HD and UHD video streaming. In: Ninth international conference on quality of multimedia experience (QoMEX), Erfurt. IEEE, pp 1–6

Lamm K, Mandl T, Womser-Hacker C, Greve W (2010a) The influence of expectation and system performance on user satisfaction with retrieval systems. In: EVIA@ NTCIR, pp 60–68

Lamm K, Mandl T, Womser-Hacker C, Greve W (2010b) User experiments with search services: methodological challenges for measuring the perceived quality. In: Proceedings of the 3rd workshop on perceptual quality of systems (PQS). ISCA/DEGA, pp 64–69

Lick E, König B, Kpossa MR, Buller V (2017) Sensory expectations generated by colours of red wine labels. J Retail Consum Serv 37:146–158

Masson J, Aurier P, d'hauteville F (2008) Effects of non-sensory cues on perceived quality: the case of low-alcohol wine. Int J Wine Bus Res 20(3):215–229

Moon S-E, Lee J-S (2015a) Perceptual experience analysis for tone-mapped HDR videos based on EEG and peripheral physiological signals. IEEE Trans Auton Ment Dev 7(3):236–247

Moon S-E, Lee J-S (2015b) EEG connectivity analysis in perception of tone-mapped high dynamic range videos. In: Proceedings of the 23rd ACM international conference on multimedia. ACM, pp 987–990

Najaf-Zadeh H, Budagavi M, Faramarzi E (2017) VR+HDR: a system for view-dependent rendering of HDR video in virtual reality. In: IEEE international conference on image processing (ICIP), pp 1032–1036

Narwaria M, Da Silva MP, Le Callet P, Pepion R (2014) Tone mapping based HDR compression: does it affect visual experience? Signal Process Image Commun 29(2):257–273

Narwaria M, Da Silva MP, Le Callet P (2015a) High dynamic range visual quality of experience measurement: challenges and perspectives. Visual signal quality assessment. Springer, Cham, pp 129–155

Narwaria M, Da Silva MP, Le Callet P (2015b) HDR-VQM: an objective quality measure for high dynamic range video. Signal Process Image Commun 35:46–60

Narwaria M, Da Silva MP, Le Callet P, Valenzise G, De Simone F, Dufaux F (2016) Quality of experience and HDR: concepts and how to measure it. High dynamic range video. Elsevier, Amsterdam, pp 431–454

Rieh SY, Belkin NJ (1998) Understanding judgment of information quality and cognitive authority in the www. In: Proceedings of the 61st annual meeting of the American Society for Information Science (ASIS), vol 35, pp 279–289

Rieh SY, Belkin N (2000) Interaction on the web: Scholars' judgement of information quality and cognitive authority. In: Proceedings of the 63rd annual meeting of the American Society for Information Science (ASIS), vol 37, pp 25–38

Sackl A, Masuch K, Egger S, Schatz R (2012a) Wireless vs. wireline shootout: how user expectations influence quality of experience. In: Fourth international workshop on quality of multimedia experience (QoMEX). IEEE, pp 148–149

Sackl A, Zwickl P, Egger S, Reichl P (2012b) The role of cognitive dissonance for QoE evaluation of multimedia services. In: IEEE Globecom workshops, pp 1352–1356

Staelens N, Moens S, Van den Broeck W, Marien I, Vermeulen B, Lambert P, Van de Walle R, Demeester P (2010) Assessing quality of experience of IPTV and video on demand services in real-life environments. IEEE Trans Broadcast 56(4):458–466

Usman MA, Usman MR, Shin SY (2015) The impact of temporal impairment on quality of experience (QoE) in video streaming: a no reference (NR) subjective and objective study. Int J Comput Electr Autom Control Inf Eng 9(8):1570–1577

van Kester S, Xiao T, Kooij RE, Brunnström K, Ahmed OK (2011) Estimating the impact of single and multiple freezes on video quality. In: Human vision and electronic imaging XVI, vol 7865. SPIE, pp 1–10

Vanderhaegen F (2017) Towards increased systems resilience: new challenges based on dissonance control for human reliability in cyber-physical&human systems. Annu Rev Control 44:316–322

Verbeke W, Viaene J (1999) Consumer attitude to beef quality labeling and associations with beef quality labels. J Int Food Agribus Mark 10(3):45–65

Wang P, Sang X, Zhu Y, Xie S, Chen D, Guo N, Yu C (2017) Image quality improvement of multi-projection 3D display through tone mapping based optimization. Opt Express 25(17):20 894–20 910

Yang X, Zhang L, Wong T-T, Heng P-A (2012) Binocular tone mapping. ACM Trans Graph 31(4):1–10

Yu P, Liu F, Geng Y, Li W, Qiu X (2015) An objective multi-layer QoE evaluation for TCP video streaming. In: IFIP/IEEE international symposium on integrated network management (IM), pp 1255–1260