

An Experimental Study on the Perceived Quality of Natively Graded versus Inverse Tone Mapped High Dynamic Range Video Content on Television

Gonzalo Luzardo · Tine Vyvey ·
Jan Aelterman · Tom Paridaens ·
Glenn Van Wallendael · Peter Lambert ·
Sven Rousseaux · Hiep Luong ·
Wouter Durnez · Jan Van Looy ·
Wilfried Philips · Daniel Ochoa

Received: date / Accepted: date

Abstract High Dynamic Range (HDR) television promises to display higher brightness and deeper black levels and thus more vivid and realistic images. However, home video distribution and video broadcasting were historically designed for what we now call standard dynamic range screens (SDR). In order to display SDR content on an HDR screen, it is explicitly or implicitly converted, in a process called inverse tone mapping (iTMO). This paper's goal is to assess the perceived quality of converted SDR content in comparison to natively graded HDR content. In doing so, this paper aims to enable content creators/distributors to make informed choices between creating/broadcasting HDR content or relying on conversion. To this end, a psychophysical experiment was performed to test how viewers evaluate the difference between natively graded HDR and a set of SDR

G. Luzardo, J. Aelterman, H. Luong, W. Philips
imec-IPI-UGent, Ghent University,
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
E-mail: {GonzaloRaimundo.LuzardoMorocho, Jan.Aelterman, Hiep.Luong, Wilfried.Philips}@UGent.be

T. Vivey, W. Durnez, J. Van Looy
imec-mict-UGent, Ghent University,
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
E-mail: Wouter.Durnez@UGent.be

T. Paridaens, G. Van Wallendael, P. Lambert
Department of Electronics and Information Systems, Ghent University,
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
E-mail: {Tom.Paridaens, Glenn.Vanwallendael, Peter.Lambert}@UGent.be

S. Rousseaux
Vlaamse Radio -en Televisieomroeporganisatie,
Auguste Reyerslaan 52, Brussels, Belgium
E-mail: sven.rousseau@vrt.be

G. Luzardo, D. Ochoa
Facultad de Ingeniería en Electricidad y Computación, ESPOL Polytechnic University,
Campus Gustavo Galindo Km. 30.5 Vía Perimetral, Guayaquil, Ecuador
E-mail: {gluzardo, dochoa}@espol.edu.ec

to HDR conversion options in a television setup. Results indicate that viewers prefer natively graded HDR content, followed by inverse tone mapping algorithms starting from videos with a compressed dynamic range. When comparing conversion options, users clearly prefer conversion from 'compressed dynamic range' SDR over 'clipped dynamic range' SDR. Users disliked videos that were naively stretched from standard SDR. In addition, a significant effect of type of sequence was found, with a preference for light scenes with low contrast.

Keywords High Dynamic Range · Inverse Tone Mapping · Subjective study

1 Introduction

High Dynamic Range (HDR) is considered an important next step in the evolution of television technology. HDR imaging overcomes the dynamic range limitations of traditional imaging by capturing the full range of the visible light spectrum (e.g., highlights and deeper tones) and colors that exist in the real world by performing operations at high bit-depths [7]. Dynamic range refers to the ratio between the brightest whites and darkest blacks present in an image, and it is commonly measured in f-stops (or simply stops), which is the logarithm of the ratio. The following definitions of the different types of Dynamic Ranges, based on the number of f-stops (the term f-stop refers to the contrast ratios), were adopted: Standard Dynamic Range (SDR) is ≤ 10 f-stops (also known as Low Dynamic Range (LDR)), Enhanced Dynamic Range (XDR) is > 10 f-stops and ≤ 16 f-stops, High Dynamic Range (HDR) is > 16 f-stops [28].

HDR technology is capable of enhancing the quality of television experience with a dynamic range comparable to the Human Visual System (HVS) [49]. Additionally, due to its truthful representation of the real world with more details and information about the scenes, it is becoming more relevant in other fields such as video game developing, medical imaging, computer vision, scientific visualization, surveillance, among others [15].

In the real world, people are capable of perceiving daylight levels from 10^1 to 10^8 candela per square meter - cd/m^2 (photopic vision), to night's luminance from approximately 10^{-1} to 10^{-6} cd/m^2 (scotopic vision) [25, 35]. The subjective television study of [10] found that the average person has a luminance dynamic range of $0.005 \text{ cd}/\text{m}^2$ to $3000 \text{ cd}/\text{m}^2$. Another small-screen viewer study [9] found that the preferred maximum luminance on average is $20000 \text{ cd}/\text{m}^2$. HDR prototype monitors are now able to display a contrast ratio of 1000000:1 with a peak luminance of $6000 \text{ cd}/\text{m}^2$. These values are much higher than the peak brightness of standard television displays, which is around $100 \text{ cd}/\text{m}^2$ [17]. Similarly, so-called "HDR compatible TVs" that can produce brightness levels up to $1000 \text{ cd}/\text{m}^2$ or 1000 nits (a "nit" is another way to describe a brightness of $1 \text{ cd}/\text{m}^2$) and a wider color palette, have become more affordable in recent years. Consequently, it is crucial to develop efficient ways of handling HDR content, including developing efficient HDR video compression algorithms and ways of dealing with legacy material, both content, and technology.

The complexity of HDR requires upgrading the entire end-to-end pipeline: Capturing, Manipulation, and Displaying. The process of capturing HDR images has been the focus of work by many researchers, artists, and photographers. Imaging

technology has advanced in such a way that capture, encoding, storage and delivery of a broad dynamic range are now possible [30, 44]. The introduction of HDR technology has provided new opportunities for content creators to enhance their storytelling and develop their creative vision [40].

Unfortunately, SDR content is still the standard in broadcasting and for regular home video distribution. Moreover, a large amount of legacy content has been recorded and/or graded in SDR. To properly display SDR content on an HDR display, an inverse tone mapping operator (iTMO) is required (i.e. SDR-to-HDR conversion). Inverse tone mapping algorithms work on reproducing real-world appearance images by using SDR images as input [38, 43]. The emergence of such SDR-to-HDR conversion brings up a question: What is the perceptual quality advantage of natively creating and distributing HDR content over using SDR-to-HDR conversion in a traditional pipeline?

To address this question, this paper presents a psychophysical experiment evaluating the perceptual quality of several approaches of inverse tone mapping and different types of video sequences with different brightness and contrast levels and comparing to natively graded HDR. We decided to carry out this study because psychophysical experimentation is considered the most reliable and accurate way to measure perceived multimedia quality and perform a quantitative comparison [26, 47]. Therefore, it gave us the opportunity to perform a reliable comparison of the results obtained by the different inverse tone mapping approaches. More particularly, we tested the following three different approaches for dynamic range expansion:

- XDR_1 : Inverse tone mapping using simple linear stretching. To obtain the input SDR content, the dynamic range of the raw video sequence is **clipped** prior to grading.
- XDR_2 : Inverse tone mapping using non-linear expansion operators. To obtain the input SDR content, the dynamic range of the raw video sequence is **clipped** prior to grading.
- XDR_3 : Inverse tone mapping using non-linear expansion operators. To obtain the input SDR content, the dynamic range of the raw video sequence is **compressed** during the grading.

Figure 1 depicts the three different approaches for dynamic range expansion we assessed. As can be seen, for XDR_1 and XDR_2 we used the same SDR content, which was obtained by clipping the dynamic range of the raw video sequence before grading. Instead, for XDR_3 the SDR content was obtained by compressing the raw video sequence during the grading. Table 1 shows visual differences between the content that was clipped prior to grade and the content where the dynamic range was compressed during the grading. Details in bright areas can be perceived in the content that was compressed during grading.

As inverse tone mapping expansion operator, we used a novel fully-automatic inverse tone mapping algorithm that is specifically designed to expand SDR video streams to HDR based on mid-level mapping [29]. This algorithm extracts features that represent the contrast, overall brightness and percentage of overexposed pixels in a frame of video and adapts the expansion curve used for the expansion of this frame to reach the most adequate mid-level (overall brightness) on the resulting HDR image. This algorithm allows expanding LDR images to the HDR domain with a peak brightness higher than 1000 nits.

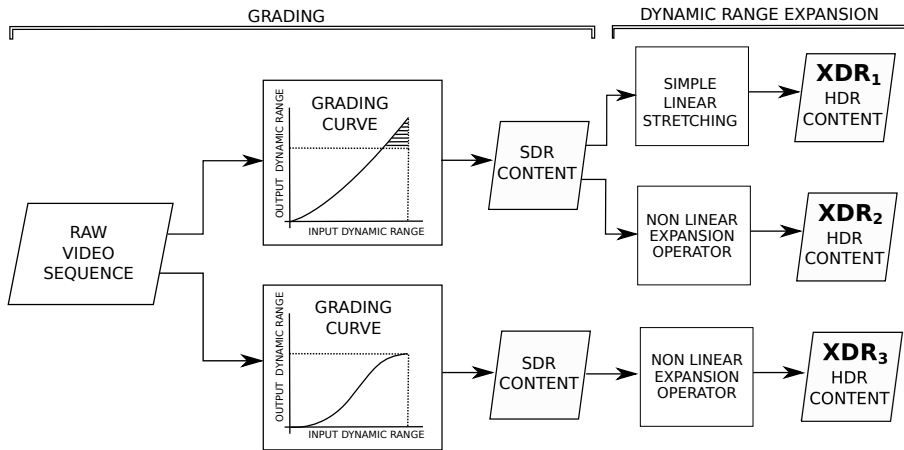




Fig. 1 Inverse tone mapping approaches.

Table 1 Visual differences between the SDR content where the dynamic range of the raw video sequence was clipped prior to grading (left) with the SDR content where the dynamic range of the raw video sequence was compressed during grading (right).

SDR content obtained by clipping the raw video sequence prior to grading (used to compute XDR ₁ and XDR ₂)	SDR content obtained by compressing the raw video sequence during grading (used to compute XDR ₃)
	

This paper is organized as follows: An overview of related works is given in Section 2. Then, motivation and research questions are detailed in Section 3. Section 4 describes the method used in our study. Results are presented in Section 5. Finally, discussions and conclusion are presented in Section 6 and Section 7, respectively.

2 Related work

2.1 Inverse Tone Mapping

The expansion of the dynamic range of standard dynamic range content can be done in different ways. From a broader viewpoint, techniques for inverse tone mapping can be divided into a) techniques based on simple linear stretching and b) more complex techniques, based on non-linear expansion operators. The first

set of techniques stretches the SDR content to HDR using a bit-depth extension technique to increase its bit-depth, commonly from 8 to 10 bits [11, 12]. The second set performs the expansion operation using a so-called inverse/reverse tone mapping operator (iTMO) and it can be classified on the basis of their operation on the image [38]. The main task of an iTMO is to recreate HDR content from input SDR content. Most of the iTMOs seek to preserve aspects of the original SDR image such as contrast, brightness, and color [6, 23, 33, 43]. Nonetheless, more advanced techniques are focused on trying to reconstruct details that might be lost on under/over-saturated regions in the SDR image [27, 31].

2.2 Subjective evaluation of Inverse Tone Mapping techniques

Inverse tone mapping algorithms and the objective evaluation of its performance have been meticulously explored. Objective evaluation techniques are usually performed by using full-reference quality metrics, where the input SDR image or HDR image is used as a reference [3, 48]; or by using a blind quality assessment approach, where the evaluation is done without information from a reference image [8].

In order to obtain a reliable rating for the quality of HDR images, researchers should take the psychophysical processes of the Human Vision System (HVS) and the subjective perception into account [46]. A previous psychophysical study has assessed how the inverse tone mapping operators perform across a wide range of exposure levels by using a criterion of subjective fidelity, asking participants how accurately the scene in the assessed HDR image is with a scene in the real world [32]. However, other studies have evaluated more specific variables such as the overall similarity, the similarity in bright and dark areas, and the color fidelity [2, 4]. A more recent study was conducted to test the veracity of the assumption that HDR video would be preferable by the end-user over SDR video [34].

3 Motivation

When making the move towards SDR television on HDR displays, it is important to conduct a psychophysical user experiment to understand the effect of different illumination and contrast levels in the video as well as the difference between different types of dynamic range expansion techniques and natively graded HDR content. Hence, the goal of the present paper is to investigate how viewers assess the quality of the produced videos with the different techniques to expand the dynamic range of SDR content to display it on HDR devices. Therefore, we propose the following research question:

- **RQ₁**: How do viewers assess the quality of HDR linearly stretched from SDR where the dynamic range of the input SDR is clipped prior to grading (XDR_1), inverse tone mapped using a non-linear expansion operator where dynamic range of the input SDR is clipped prior to grading (XDR_2), and inverse tone mapped using a non-linear expansion operator where dynamic range is compressed during grading (XDR_3)? With this research question, we will specifically focus on the following questions:
 - **RQ_{1.1}** Is it useful to natively (manually) grade HDR (HDR vs. XDR_1 , XDR_2 , XDR_3)?

- **RQ_{1.2}** Is it useful to apply an inverse tone mapping algorithm when upscaling from SDR to HDR, or could we just as well stretch linearly (XDR₁ vs. XDR₂, XDR₃)?
- **RQ_{1.3}** Is it useful, when grading for SDR and knowing that it will be expanded to HDR, to reduce the range during grading as opposed to before (XDR₂ vs. XDR₃)?

Moreover, we want to know if the type of displayed video sequence influences viewers’ assessment of the quality of the video. Therefore, we propose the second research question.

- **RQ₂**: Does the nature of the scene (light vs. dark, large vs. small contrast) affect how viewers assess the quality of the video?

To answer these questions, we set up psycho-visual experiments that are described in the following sections.

4 Method

4.1 Participants

Before the actual experiment, a pilot study of the television experiment ($N = 11$, $M_{age} = 29.64$, $SD_{age} = 5.24$, 73% male, 24 - 39 years old) was conducted. Forty-three people ($M_{age} = 26.44$, $SD_{age} = 6.3$, 49% male) participated in the main within-subjects experiment, of which 32 were non-experts users and 11 were experts in video and image processing.

4.2 Design

A pilot study was conducted to: 1) evaluate the efficacy of the study design, 2) to evaluate the analytical methodology for measuring the perceived quality of the HDR video content, and 3) to have an initial indication of the scores of perceived video quality and the effect of the independent variables *video sequence* and *dynamic range expansion approach*.

In the effective experiment, we examined the subjective perception of the quality of a series of video sequences. To achieve this, we used an AB design with reference, based on the double-stimulus impairment scale (DSIS) [19]. More specifically, subjects assessed overall quality, vividness of color, and contrast quality differences between two videos, using the first video as a reference to rate the second video.

The overall quality refers to the human perceptual judgment of overall quality of the second video with respect to the first one. The subjective criteria of vividness of colors, also known as colorfulness, refers to how saturated-brilliant is the color appearance of the colored objects in the video, vague and dim (low quality) or perfectly clear and vivid (high quality) as real seeing [22]. And, the contrast refers to how the difference between blacks and whites are perceived by the observer, the video do not exhibit a great deal of difference between its lights and darks and appears flat or dull (low quality) or they exhibit a full range of tones from black to white, with dark shadows and bright highlights [39].

In each trial, the same video sequence was shown twice, with a (possible) different grading method. The experiment was fully counterbalanced, implying that every dynamic range expansion technique (XDR_1 , XDR_2 , XDR_3) and the natively graded HDR, was compared with every dynamic range expansion technique (including the natively graded HDR) twice (once with video 1 as reference stimulus and video 2 as test stimulus, and once in the other way). There were 6 different video sequences, amounting to a total of 96 trials (4 reference grading methods x 4 test grading methods x 6 video sequences).

4.3 Materials

4.3.1 Professionally graded video content

Professionally graded video content refers to video sequences that have been obtained by altering and/or enhancing a master (or a raw input video), in order to be properly displayed on an specific display. This process involves an artistic step, where the "grader" manipulates the input to better express the director's artistic intentions on the final graded content, which is extremely important to offer a reliable base to the discussion on how the content is perceived by the viewers.

The process for mastering motion pictures usually consists of creating a theatrical master and then applying a so-called "trim pass" to create a home video master, either in HDR or SDR [40]. This process can be done simultaneously by dual-mastering, which means to have two different pipelines, one to create an LDR master (natively graded on a SDR display) and another to create an HDR master (natively graded on an HDR display); or by grading the HDR master first and then obtain the SDR version by tone mapping [37]. Although the SDR master has a narrower range of luminance and color gamut than the HDR master, both have been graded to transmit the same artistic intentions, as well as, to ensure the best consumer experience when they are displayed. Indisputably, the HDR domain provides more room for filmmakers to improve their storytelling in comparison with SDR domain. However, in the end, their artistic intentions must be reflected in both types of content.

4.3.2 Video sequences used in our study





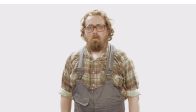

In this work we analyzed if natively graded HDR video (content graded using a HDR display as a reference), linearly stretched SDR (XDR_1), and two types of inverse-tone mapped HDR sequences (according to the input SDR content) using non-linear expansion operators (XDR_2 and XDR_3), elicit different viewer responses on the assessment of HDR video quality. Additionally, we investigated whether the nature of the scene influenced the perceived HDR video quality.

For this research, the Flemish Radio and Television Broadcasting Organization (Vlaamse Radio- en Televisieomroeporganisatie-VRT) produced a short film of about 5 minutes which contained a wide variety of scenes in different luminance, contrast, and transition conditions. This short film was professionally graded by experts at VRT in HDR and SDR (both, by clipping the dynamic range prior grading and by compressing the dynamic range during grading) by dual-grading, using a SIM2-HDR47E and a Barco type RHDM-2301 as reference displays, respectively.

The SIM2 HDR display contains 2202 individually modulated LED back-lights, which can reach up to 6000 nits of peak brightness [1]. Both versions, the HDR and SDR, were produced in full HD (1920x1080) at 24 fps. For the experiment, we selected six different video sequences of 15 seconds each, which represent all different scenes present in the short film. The selected sequences are described in Table 2.

All video sequences used in this study contain a low intensity of motion activity. The motion activity is defined as the perceived subjective degree of activity, or amount of motion, in a video sequence [14, 21].

Table 2 Video sequences used in the study

Thumbnail	Scene name	Label	Description
	Average	AVG	An <i>average</i> scene filmed outdoors during the day with moderate brightness
	Street	DHC	A <i>dark</i> scene filmed at night with <i>high contrast</i>
	Mannequin	DLC	A <i>dark scene</i> filmed with dim light with <i>low contrast</i>
	Welding	LHC	A scene filmed with dim light that contains <i>bright</i> parts with <i>high contrast</i>
	Heaven	LLC	A <i>bright</i> scene with high luminance and <i>low contrast</i>
	Garage	TRA	A scene with a <i>transition</i> from dark to bright

























The short film, both HDR and SDR version, and sequences used in this study are available at: http://telin.ugent.be/~gluzardo/experimental_study

4.4 Procedure

First, participants sat in a waiting room close to the experiment room to adapt to the ambient light of 50 lux. During the adaptation, participants read a document with information about the experiment, including the explanation about the three

subjective perception of quality aspects that they were going to assess for each video sequence: overall quality, vividness of colors, and contrast. This document included a detailed description of these three aspects and examples of images with low and high quality scores for each one. Then, after adapting for five minutes, participants were seated in the experiment room with a viewing distance of 1 meter from the HDR screen. The participants had to go through a number of trials in which first a reference video was presented with a duration of 15 seconds, followed by a completely gray screen with a fixation cross positioned in the middle of the screen for a duration of two seconds. Next, participants were asked to rate the general quality, the quality of the colors, and the quality of the contrast in the second video. All three ratings were indicated on a 7-point Likert scale for (-3: much worse, -2: worse, -1: slightly worse, 0: the same, +1: slightly better, +2: better, +3: much better) using the left mouse button. Clicking a rating also started the next trial. All information for the participants was projected in the center of the television screen while maintaining a black background level. In this way, the attention of the participants remained on the screen. Table 3 shows a preview of the sequences that participants had to rate during this experiment. Images have been tone mapped in order to properly show them in this document.

Table 3 Preview of the sequences that participants had to rate during this experiment. *Images have been tone mapped in order to properly show them in this document.*

HDR	XDR ₁	XDR ₂	XDR ₃
			
			
			
			
			
			

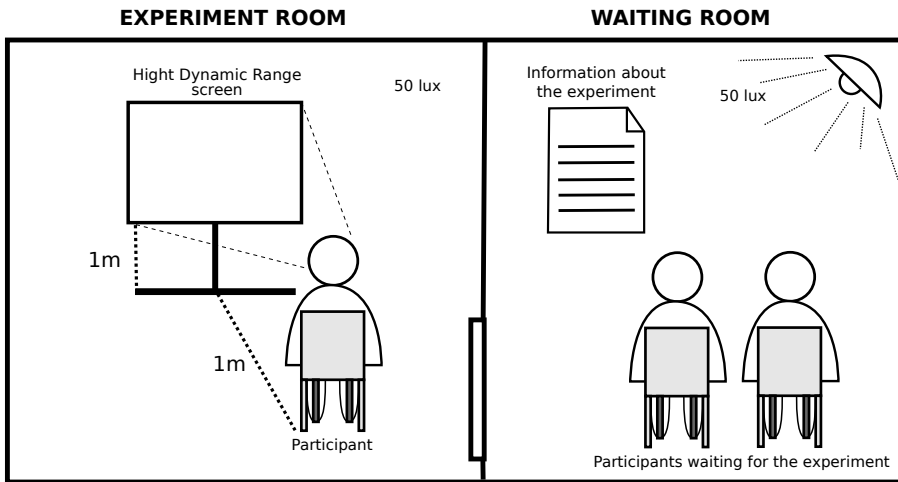


Fig. 2 Experimental setup.

The experiment was divided into two sets spread over two days, with each set lasting around 40 minutes. The sequence of conditions was randomized. To ensure the accuracy of the answers, the experiments were counterbalanced. Because of this procedure, and to avoid interaction between the different participants, the tests were taken individually so only one participant was present in the experiment room at any given time. The ambient light of the experiment room was 50 lux, to simulate the ambient light of a dark living room. Figure 2 shows the setup of the experiment.

4.5 Data-analysis

The data in this experiment are the responses to three presented Likert scales, gauging for perceptual differences in general quality, color quality, and contrast quality. The response options on Likert scales have a rank order, but cannot – strictly speaking – be considered equidistant, as the distance between the values 'much worse' and 'worse' may not mean the same as that between 'same' and 'slightly better' [24]. In other words, Likert scales yield ordinal data, not interval data. Still, such data is frequently treated as if it were normally distributed and analyzed using parametric statistical tests. This practice has been (and still is) the subject of some academic debate, the full extent of which is out of the scope of this paper [20, 24, 36, 42]. In this study, we used parametric tests. Aside from the theoretical arguments in favor of the robustness of parametric tests for Likert data [36, 42], we also contend that the symmetrical nature of both the Likert scales and the experimental design precludes the emergence of skewed data - a problem typical of Likert response scale data [36] (see fig. 3 for response distributions). In addition, our scales also showed numerical labels (e.g., '-2') along with the semantic labels (e.g., 'worse'), further deflating potential criticism.

We analyzed the response data using linear mixed-effect models with a Gaussian link function, as implemented in the R package 'lme4' [5]. Mixed-effect models

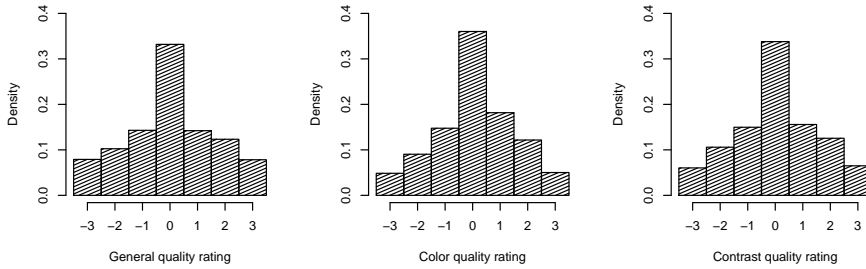


Fig. 3 Quality rating histograms.

can take individual differences into account (e.g., different baselines in observers' response tendency), and allow us to more accurately estimate the coefficients of the fixed terms of our statistical model (i.e., the experimental effects on a population level) [45].

Statistical modeling consisted of the following steps. First, each set of responses (general, color and contrast quality ratings) was fit with a model containing *dynamic range expansion approach* (XDR₁, XDR₂, XDR₃, or HDR) and the interaction between *dynamic range expansion approach* and *video sequence* (AVG, DHC, DLC, LHC, LLC, and TRA) as fixed factors. The *dynamic range expansion approach* variable represents the technique used for dynamic range expansion (including natively graded HDR) of the sequence that was judged (i.e., the second sequence of the double-stimulus evaluation). We purposely did not include the *dynamic range expansion approach* of the reference sequence in our model, because (1) this complicates the analyses without gaining useful information, and (2) since all types of approaches were compared with every other type (including itself) an equal amount of times, we are still left with a fair comparison. We also left out the factor *video sequence* – outside of its interaction with the variable *dynamic range expansion approach* – as the main effect of this variable would be nonsensical: each rating represents a comparison of different dynamic range expansion techniques, yet applied to the same video sequence. Any effect on the observer's responses should therefore not be attributed to the main effect of the sequence itself. By default, a random intercept was added conditional on the *subject* variable. We then assessed whether it was appropriate to add a random effect for either of the within-subject variables, by checking whether such addition significantly increased the model's goodness of fit.

In a second step we sought out the most parsimonious model by systemically restricting the full model (including all relevant factors and interactions), and comparing the goodness of fit using likelihood-ratio tests. In a third and final step, we inspected the analysis of variance table and tested specific hypotheses. For a similar approach, see [13, 16, 41]. As we test an increasing number of hypotheses, we also increase the risk of encountering a Type I error - a false positive. To mitigate this risk, a correction is typically applied. In this study, we adhered to the Holm-Bonferroni approach [18], as reflected in the corrected p-values (p_{corr}). The significance level used was $\alpha = .05$.

5 Results

The model that best fit the data included a random subject-based intercept, as well as a random effect for the *dynamic range expansion approach* variable. In this model, we detected a significant interaction effect (*dynamic range expansion approach* x *video sequence*), removing the need for further model restriction. Importantly, yet unsurprisingly, this was the case for all ratings – general quality, color quality, and contrast quality – leaving us with the following overall statistical model (using lme4 [5] notation):

$$Q \sim DREA + (DREA:SEQ) + (1 + DREA | PP)$$

In this formula, Q stands for the quality rating, DREA for *dynamic range expansion approach*, SEQ for *Video Sequence*, and PP for *participant*. All hypotheses were tested using this model.

5.1 Research Question 1 (RQ₁)

Our first research question pertained to the impact of different dynamic range expansion approaches on viewers' perception of the general quality, color quality, and quality of contrast. This research question was split up into three parts, which were answered by testing the corresponding contrasts (see Table 4).

RQ_{1.1}: Is it useful to natively grade HDR (HDR vs XDR₁, XDR₂, XDR₃)? We found that natively graded HDR content indeed predicted better ratings, both for general quality ($\chi^2 = 338.830$, $p < .001$), color quality ($\chi^2 = 93.425$, $p < .001$), and contrast quality ($\chi^2 = 123.85$, $p < .001$).

RQ_{1.2}: Is it useful to apply an inverse tone mapping algorithm rather than linearly stretch the range (XDR₁ vs XDR₂, XDR₃)? We found that the use of an inverse tone mapping algorithm predicted superior quality ratings across the board (general quality: $\chi^2 = 1314.000$, $p < .001$; color quality: $\chi^2 = 176.560$, $p < .001$; contrast quality: $\chi^2 = 265.250$, $p < .001$).

RQ_{1.3}: Is it useful to compress range during grading, rather than clipped it beforehand (XDR₃ vs XDR₂)? While compression during grading predicted better general quality ratings ($\chi^2 = 7.984$, $p = .005$), and better color quality ratings ($\chi^2 = 10.165$, $p = .001$), it was not associated with higher ratings of contrast quality ($\chi^2 = 3.6328$, $p = .057$).

Table 4 Contrast table for Research Question 1 (RQ₁)

RQ	General			Color			Contrast		
	1.1	1.2	1.3	1.1	1.2	1.3	1.1	1.2	1.3
χ^2	338.83	1314	7.984	93.425	176.56	10.165	123.85	265.25	36.328
p	0.001	0.001	0.005	0.001	0.001	0.001	0.001	0.001	0.057
p_{corr}	0.021	0.021	0.035	0.021	0.021	0.021	0.021	0.021	0.399

Significant p -values indicated by **boldface** ($\alpha = .05$).

p_{corr} rows indicate corrected p -values according to Holm-Bonferroni method.

5.2 Research Question 2 (RQ₂)

Our second research question pertained to the characteristics of the video sequence that was used in the comparison, and whether they influenced quality judgments. The significant interaction effect (*dynamic range expansion approach x video sequence*) in our model suggests that this is indeed the case. To further analyze the nature of this interaction, we tested all sub-hypotheses of RQ₁ again, now applied to each separate video sequence (labeled as RQ_{2.1}, RQ_{2.2}, and RQ_{2.3}). This resulted in a (corrected) set of 54 (3 x 3 x 6) contrasts, which are shown in Table 5 (see also Fig. 4).

Results for RQ_{2.1} indicated that natively graded HDR content received better quality ratings compared to the combined ratings of XDR₁, XDR₂, and XDR₃, regardless of the video sequence involved (all $p_{corr} < .05$), barring two exceptions. Color quality ratings for the DLC were only marginally significant after correction ($p_{corr} = .054$), with better ratings for HDR sequences. In addition, contrast quality ratings for this sequence were not significantly different between the tested groups ($p_{corr} > .05$).

Visual inspection of the data suggested, however, that these effects are driven to no small degree by the low ratings given in the XDR₁ condition. As XDR₁ was expected to perform worst in the ratings, we ran post-hoc tests comparing HDR with XDR₂ and XDR₃ – this time barring XDR₁. All contrasts were significant ($p_{corr} < .05$), i.e. HDR outperformed XDR₂ and XDR₃ for all sequences and in all ratings, except for the following conditions. Our test did not reach significance for any quality ratings awarded to DLC video sequences (all $p_{corr} > .05$), and in LHC video sequences the contrast was not significant for color rating data ($\chi^2 = 1.575$, $p_{corr} = .210$).

With respect to RQ_{2.2}, we found that HDR content obtained by using an inverse tone mapping algorithm led to significantly higher ratings of general, color and contrast quality across all video sequences (all $p_{corr} < .05$). Finally, results were a little more complex for RQ_{2.3}, as the applied correction method significantly impacted the obtained results. More specifically, 6 contrasts were initially significant, predicting a difference in quality rating between XDR₂ and XDR₃ in the following conditions: AVG, DHC, and LHC for general quality ratings, AVG and LHC for color quality ratings, and DHC for contrast quality ratings (all $p_{corr} < .05$). After correction, however, only the color quality rating for the AVG video sequence tested significant, predicting a better rating for XDR₃ compared to XDR₂.

6 Discussion

In this paper, we assessed the perceptual quality of the most common techniques of inverse tone mapping using different types of video sequences, from dark to bright scenes and from low to high contrast, in comparison to natively graded HDR. We first analyzed how observers perceived video sequences in terms of general, color, and contrast quality, depending on the applied dynamic range expansion approach. Second, we investigated whether the visual characteristics of the video sequence affect an observer's perception when comparing these approaches.

Our analyses clearly demonstrate the impact of different dynamic range expansion approaches on perceived general image quality, image color quality, and

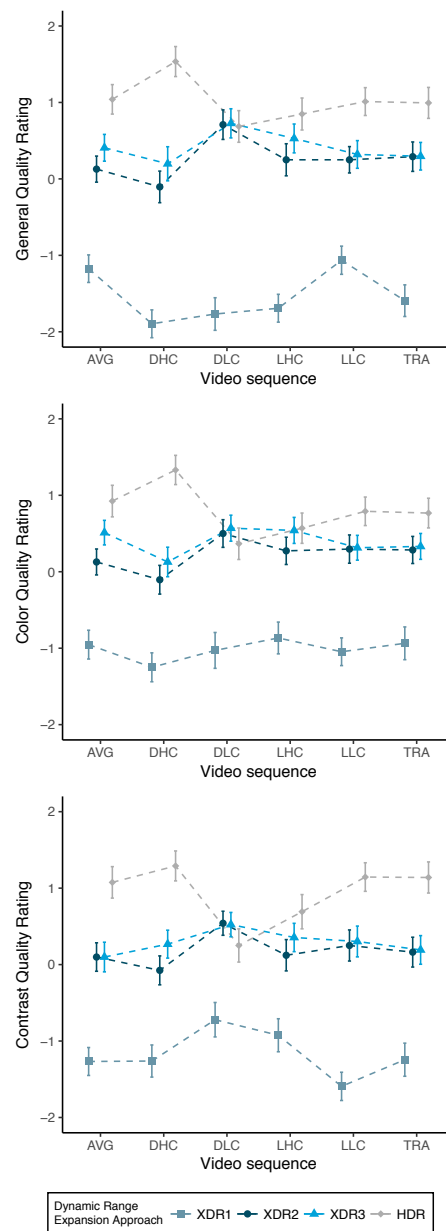


Fig. 4 Average quality ratings for video sequences subjected to different dynamic range expansion approaches, including 95% confidence intervals.

Table 5 Contrast table for Research Question 2 (RQ₂)

Video se- quence	Quality Rating									
		General			Color			Contrast		
	RQ	2.1	2.2	2.3	2.1	2.2	2.3	2.1	2.2	2.3
AVG	χ^2	101.600	144.600	4.141	57.583	79.128	8.750	96.277	97.728	0
	p	.001	.001	.042	.001	.001	.003	.001	.001	1
	p_{corr}	.021	.021	.210	.021	.021	.021	.021	.021	1
DHC	χ^2	294.750	262.270	4.860	164.690	77.689	3.214	127.370	96.484	6.302
	p	.001	.001	.027	.001	.001	.073	.001	.001	.012
	p_{corr}	.021	.021	.162	.021	.021	.292	.021	.021	.096
DLC	χ^2	4.999	429.670	.016	6.765	119.380	.289	.864	82.182	.016
	p	.001	.001	.899	.009	.001	.591	.353	.001	.898
	p_{corr}	.021	.021	1	.054	.021	1	1	.021	1
LHC	χ^2	85.926	301.320	4.141	18.750	79.128	4.250	33.206	7.785	2.897
	p	.001	.001	.042	.001	.001	.039	.001	.001	.089
	p_{corr}	.021	.021	.210	.021	.021	.195	.021	.021	.534
LLC	χ^2	89.427	126.540	.259	47.643	89.184	.018	104.520	182.910	.147
	p	.001	.001	.611	.001	.001	.893	.001	.001	.702
	p_{corr}	.021	.021	1	.021	.021	1	.021	.021	1
TRA	χ^2	114.220	247.570	.002	41.539	75.556	.129	96.799	105.799	.045
	p	.001	.001	.966	.001	.001	.720	.001	.001	.831
	p_{corr}	.021	.021	1	.021	.021	1	.021	.021	1

Significant p -values indicated by **boldface** ($\alpha = .05$).

p_{corr} rows indicate corrected p -values according to Holm-Bonferroni method.

image contrast quality (RQ₁). We found strong evidence for the added value of natively/manually graded HDR footage, as these sequences were perceived to be of higher quality in each of the measured dimensions. As expected, we found that a more complex inverse tone mapping algorithm outperforms techniques based on simple linear stretching in all quality ratings. Results also show that compression of the high dynamic range during grading is useful, as it leads to both better ratings of general and color quality compared to compression after grading. However, it does not appear to yield a quality gain with regard to contrast quality perception.

Follow-up analyses were conducted to assess whether the characteristics of the video sequence itself can influence these effects (RQ₂). Our results suggest that this is indeed the case. As expected, manual grading of HDR outperforms the other expansion approaches for every type of video sequence, bar one: the dark, low contrast fragment did not yield better contrast quality ratings, and only marginally better color quality ratings. Since visual inspection of our data suggested that these results were primarily driven by the low ratings granted to linearly stretched HDR images (XDR₁), we then exclusively compared HDR to inverse tone mapping algorithms (XDR₂ and XDR₃). Again, we found that natively graded content outperforms the use of an inverse tone mapping algorithm on all fronts, with two exceptions. No effects were found for the dark, low contrast fragment, nor was

color quality perceived differently in the light, high contrast sequence. The use of an inverse tone mapping algorithms is found to outperform a linear stretch approach, regardless of the visual characteristics of the video sequence. Finally, some evidence was found in favor of compressing the dynamic range during grading (XDR₃), rather than clipping it (XDR₂), specifically for the color quality of average video sequences. It is worth mentioning, however, that many of these contrasts were found to yield significant results prior to our p-value correction. Since this experiment was not designed to compare specific *video sequence* x *dynamic range expansion approach* combinations, these results were numbed by the rigorous corrections for multiple testing. Still, our findings support the practice of compressing the dynamic range during grading, potentially resulting in better overall video quality. Follow-up research is warranted.

7 Conclusion

This study yields a threefold conclusion: First, it is still worth to invest in manual HDR recording, grading, transmission and display for new, high-quality content. Our study shows a clear quality advantage, with one notable exception: footage that is low in contrast, and on the darker end of the spectrum. In that case, inverse tone mapping algorithms yield results that are similar in quality. Second, barring natively grading, there is no good reason to linearly stretch SDR content – inverse tone mapping algorithms perform better on all fronts. Third, and finally, for content being recorded using the existing SDR pipeline, our study serves as an endorsement for the practice of compressing the dynamic range during grading rather than clipping it. Our analyses show that this approach, more often than not, edges out an additional increase in perceived quality when being inversely tone-mapped to HDR.

Acknowledgements This work was supported by the imec-ICON-HD²R project, co-funded by imec, a digital research institute founded by the Flemish Government. Project partners are Barco, Grass Valley, Limecraft, VRT, and Grid. The work of G. Luzardo was supported by Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) and Escuela Superior Politécnica del Litoral (ESPOL). Jan Aelterman is currently supported by a Ghent University postdoctoral fellowship (BOF15/PDO/003).

References

1. (2017) SIM2 HDR display - HDR47ES6MB. URL <http://hdr.sim2.it/hdrproducts/hdr47es6mb>
2. Abebe MA, Pouli T, Kervec J (2015) Evaluating the color fidelity of itmos and hdr color appearance models. *ACM Transactions on Applied Perception (TAP)* 12(4):14
3. Aydin TO, Mantiuk R, Myszkowski K, Seidel HP (2008) Dynamic range independent image quality assessment. *ACM Transactions on Graphics (TOG)* 27(3):69
4. Banterle F, Ledda P, Debattista K, Bloj M, Artusi A, Chalmers A (2009) A psychophysical evaluation of inverse tone mapping techniques. In: *Computer Graphics Forum, Wiley Online Library*, vol 28 (1), pp 13–25

5. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1):1–48
6. Bist C, Cozot R, Madec G, Ducloux X (2017) Tone expansion using lighting style aesthetics. *Computers & Graphics* 62:77–86
7. Chalmers A, Karr B, Suma R, Debattista K (2016) Fifty shades of hdr. In: *Digital Media Industry & Academic Forum (DMIAF)*, IEEE, pp 53–58
8. Chen CR, Chiu CT, Chang YC (2011) Inverse tone mapping operator evaluation using blind image quality assessment. In: *Asia-Pacific Sign. and Information Proc. Association Annual Summit and Conf., APSIPA Oct*
9. Daly S, Kunkel T, Sun X, Farrell S, Crum P (2013) 41.1: Distinguished paper: Viewer preferences for shadow, diffuse, specular, and emissive luminance limits of high dynamic range displays. In: *SID Symposium Digest of Technical Papers*, Wiley Online Library, vol 44 (1), pp 563–566
10. Daly S, Kunkel T, Sun X, Farrell S, Crum P (2013) Preference limits of the visual dynamic range for ultra high quality and aesthetic conveyance. In: *Human Vision and Electronic Imaging XVIII*, International Society for Optics and Photonics, vol 8651, p 86510J
11. Daly SJ, Feng X (2003) Bit-depth extension using spatiotemporal microdither based on models of the equivalent input noise of the visual system. In: *Color Imaging VIII: Processing, Hardcopy, and Applications*, International Society for Optics and Photonics, vol 5008, pp 455–467
12. Daly SJ, Feng X (2004) Decontouring: Prevention and removal of false contour artifacts. In: *Human Vision and Electronic Imaging IX*, International Society for Optics and Photonics, vol 5292, pp 130–150
13. De Paepe AL, Crombez G, Legrain V (2017) Remapping nociceptive stimuli into a peripersonal reference frame is spatially locked to the stimulated limb. *Neuropsychologia* 101:121–131
14. Divakaran A (2001) An overview of mpeg-7 motion descriptors and their applications. In: *International Conference on Computer Analysis of Images and Patterns*, Springer, pp 29–40
15. Dong Y, Pourazad MT, Nasiopoulos P (2016) Human visual system-based saliency detection for high dynamic range content. *IEEE Transactions on Multimedia* 18(4):549–562
16. Durnez W, Van Damme S (2015) Trying to Fix a Painful Problem: The Impact of Pain Control Attempts on the Attentional Prioritization of a Threatened Body Location. *The Journal of Pain* 16(2):135–143
17. Hanhart P, Korshunov P, Ebrahimi T, Thomas Y, Hoffmann H (2015) Subjective quality evaluation of high dynamic range video and display for future tv. *SMPTE Motion Imaging Journal* 124(4):1–6
18. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* pp 65–70
19. ITU (2012) Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service. *International Telecommunication Union* 13:1–48
20. Jamieson S, et al. (2004) Likert scales: how to (ab) use them. *Medical education* 38(12):1217–1218
21. Jeannin S, Divakaran A (2001) Mpeg-7 visual motion descriptors. *IEEE Transactions on Circuits and Systems for Video Technology* 11(6):720–724

22. Khanh T, Bodrogi P, Vinh Q, Stojanovic D (2017) Colour preference, naturalness, vividness and colour quality metrics, part 1: Experiments in a room. *Lighting Research & Technology* 49(6):697–713
23. Kovaleski RP, Oliveira MM (2014) High-quality reverse tone mapping for a wide range of exposures. In: *Graphics, Patterns and Images (SIBGRAPI), 2014 27th SIBGRAPI Conference on*, IEEE, pp 49–56
24. Kuzon Jr WM, Urbanchek MG, McCabe S (1996) The seven deadly sins of statistical analysis. *Annals of plastic surgery* 37(3):265–272
25. Ledda P, Santos LP, Chalmers A (2004) A local model of eye adaptation for high dynamic range images. In: *Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, ACM, pp 151–160
26. Lee JS, De Simone F, Ebrahimi T (2011) Subjective quality evaluation via paired comparison: Application to scalable video coding. *IEEE Transactions on Multimedia* 13(5):882–893
27. Lee S, An GH, Kang SJ (2018) Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *arXiv preprint arXiv:180106277*
28. Luthra A, François E, Husak W (2016) Requirements and use cases for hdr and wcg content distribution. *ISO/IEC JTC 1/SC 29/WG 11 (MPEG) Doc N15084*
29. Luzardo G, Aelterman J, Luong H, Philips W, Ochoa D, Rousseaux S (2018) Fully-automatic inverse tone mapping preserving the content creator's artistic intentions. In: *Picture Coding Symposium 2018*, IEEE, pp 53–58
30. Mai Z, Mansour H, Nasiopoulos P, Ward RK (2013) Visually favorable tone-mapping with high compression performance in bit-depth scalable video coding. *IEEE Transactions on Multimedia* 15(7):1503–1518
31. Marnerides D, Bashford-Rogers T, Hatchett J, Debattista K (2018) Expand-net: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. *arXiv preprint arXiv:180302266*
32. Masia B, Agustin S, Fleming RW, Sorkine O, Gutierrez D (2009) Evaluation of reverse tone mapping through varying exposure conditions. *ACM transactions on graphics (TOG)* 28(5):160
33. Masia B, Serrano A, Gutierrez D (2017) Dynamic range expansion based on image statistics. *Multimedia Tools and Applications* 76(1):631–648
34. Mukherjee R, Debattista K, Bashford-Rogers T, Waterfield B, Chalmers A (2016) A study on user preference of high dynamic range over low dynamic range video. *The Visual Computer* 32(6-8):825–834
35. Myszkowski K, Mantiuk R, Krawczyk G (2008) High dynamic range video. *Synthesis Lectures on Computer Graphics and Animation* 1(1):1–158
36. Norman G (2010) Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15(5):625–632
37. Pouli T, Pines J (2016) Hdr content creation: creative and technical challenges. In: *ACM SIGGRAPH 2016 Courses*, ACM, p 14
38. Reinhard E, Heidrich W, Debevec P, Pattanaik S, Ward G, Myszkowski K (2010) High dynamic range imaging: acquisition, display, and image-based lighting. *Morgan Kaufmann*
39. Shokrollahi A, Mahmoudi-Aznavah A, Maybodi BMN (2017) Image quality assessment for contrast enhancement evaluation. *AEU-International Journal*

- of Electronics and Communications 77:61–66
40. Smith M, Zink M (2015) Managing hdr content production and display device capabilities. pp 11.–11., DOI 10.1049/ibc.2015.0031
 41. Verbruggen F, Aron AR (2010) Theta burst stimulation dissociates attention and action updating in human inferior frontal cortex. *Proceedings of the National Academy of Sciences* 107(31):13966–13971
 42. Wadgave U, Khairnar MR (2016) Parametric tests for likert scale: For and against. *Asian journal of psychiatry* 24:67–68
 43. Wang TH, Chiu CW, Wu WC, Wang JW, Lin CY, Chiu CT, Liou JJ (2015) Pseudo-multiple-exposure-based tone fusion with local region adjustment. *IEEE Transactions on Multimedia* 17(4):470–484
 44. Ward G, Simmons M (2006) Jpeg-hdr: A backwards-compatible, high dynamic range extension to jpeg. In: *ACM SIGGRAPH 2006 Courses*, ACM, p 3
 45. West BT, Welch KB, Gałecki AT (2007) *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman and Hall/CRC
 46. Winkler S (2005) *Digital video quality: vision models and metrics*. John Wiley & Sons
 47. You J, Korhonen J, Perkis A, Ebrahimi T (2011) Balancing attended and global stimuli in perceived video quality assessment. *IEEE Transactions on Multimedia* 13(6):1269–1285
 48. Zerman E, Valenzise G, Dufaux F (2017) An extensive performance evaluation of full-reference hdr image quality metrics. *Quality and User Experience* 2(1):5
 49. Zhang Y, Reinhard E, Bull D (2011) Perception-based high dynamic range video compression with optimal bit-depth transformation. In: *Image Processing (ICIP), 2011 18th IEEE International Conference on*, IEEE, pp 1321–1324