Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09532-y

Vision models fine-tuned by cinema professionals for High Dynamic Range imaging in movies

Praveen Cyriac[1] · Trevor Canham[1] · David Kane[1] · Marcelo Bertalmío[1]

Received: 30 September 2019 / Revised: 29 June 2020 / Accepted: 4 August 2020 /
Published online: 15 September 2020

Abstract

Many challenges that deal with processing of HDR material remain very much open for the film industry, whose extremely demanding quality standards are not met by existing automatic methods. Therefore, when dealing with HDR content, substantial work by very skilled technicians has to be carried out at every step of the movie production chain. Based on recent findings and models from vision science, we propose in this work effective tone mapping and inverse tone mapping algorithms for production, post-production and exhibition. These methods are automatic and real-time, and they have been both fine-tuned and validated by cinema professionals, with psychophysical tests demonstrating that the proposed algorithms outperform both the academic and industrial state-of-the-art. We believe these methods bring the field closer to having fully automated solutions for important challenges for the cinema industry that are currently solved manually or sub-optimally. Another contribution of our research is to highlight the limitations of existing image quality metrics when applied to the tone mapping problem, as none of them, including two state-of-the-art deep learning metrics for image perception, are able to predict the preferences of the observers.

Keywords High dynamic range · Vision models · Visual perception · Tone mapping · Inverse tone mapping · Cinema post-production

1 Introduction

High Dynamic Range (HDR) technology can provide a never before seen increase in contrast, colour and luminance. The stunning appearance that HDR images can have has led the industry to see HDR technology as the next generation of video content and is expected to deliver a transforming user experience, representing for some "*the most exciting format to come along since color TV*" [22]. The contribution of HDR to the sense of immersion stems from the fact that HDR images appear much more faithful to an actual perceived scene, allowing the viewer to resolve details in quite dark or quite bright regions, to distinguish

✉ Marcelo Bertalmío
marcelo.bertalmio@upf.edu

[1] Universitat Pompeu Fabra, Barcelona, Spain

subtle colour gradations, to perceive highlights as much brighter than diffuse white surfaces, all things that we associate with everyday experiences and that cannot be reproduced using SDR (Standard Dynamic Range) systems. This also allows filmmakers to overcome artistic limitations that have existed since the inception of cinema, giving them the "*perceptual tools that fine artists working in the medium of painting have had for hundreds of years*" [77] and that provide compelling new opportunities for storytelling. Besides the enthusiasm about the unparalleled improvement in picture quality for new content, the industry expects, through re-purposing for HDR, to monetize existing film and TV libraries once HDR screens are established, in theaters, at home and on mobile devices.

All the possibilities that the HDR format could offer, in terms of revenue and market growth for companies and in terms of improved user experience for the viewer, depend upon the existence of a fully functional HDR ecosystem. But we are currently a long way from such an ecosystem, as stated in recent reports from several international organizations and standardization bodies [29, 70, 75] as well as very recent publications from the motion picture and TV community [9, 24, 58, 64, 66, 78]. This is due to open challenges happening at all stages of the production chain, from capture to display. While some of these challenges are technical (e.g. developing HDR cinema projectors [9]), most of them stem from the fact that the interactions of HDR images with the human visual system are quite complex and not yet fully understood, not even in the vision science community, for a number of reasons (that may include the preference of vision studies to use synthetic instead of natural stimuli, and modeling vision as a cascade of linear and nonlinear filters, which might be too restricting: see [10] for a comprehensive review on the subject). Furthermore, cinema artists complain that the software tools they work with are based on very basic vision science; for instance, noted cinematographer Steve Yedlin says that: "*Color grading tools in their current state are simply too clunky [...]. Though color grading may seem complex given the vast number of buttons, knobs and switches on the control surface, that is only the user interface: the underlying math that the software uses to transform the image is too primitive and simple to achieve the type of rich transformations [...] that define the core of the complex perceptual look*" [87].

Among the consequences of this lack of accurate vision models we can cite the following:

- There are no guidelines on how to shoot and post-produce HDR footage in a way that maximizes the expressive capabilities of the HDR medium while considering the visual perception of HDR images.
- There is a need for colour management and grading tools so that the mastering process has a simple workflow that can deliver consistent images across the whole spectrum of possible display devices and environments.
- There is a need for high quality real-time tools for conversion from HDR to SDR formats (so-called tone mapping (TM) operators) and vice-versa (inverse tone mapping (ITM) algorithms), and among standard and wide colour gamuts, which are essential for HDR capture, monitoring, post-production, distribution and display, for cinema and TV, given that SDR and HDR displays will co-exist for the foreseeable future.

Regarding this latter point, starting from the mid 1990s the academic literature has seen a *very* significant number of works on the subject of TM (and, to a lesser degree, ITM), following diverse goals (e.g. emulating perception or improving appearance), based on diverse ideas (e.g. emulating photoreceptor response or emulating techniques by photographers), intended for still images or video sequences, and normally evaluated through user tests with regular observers or objective metrics specific for tone mapping. The literature on the topic is so vast that we have decided to concentrate just on the seminal works on vision-based tone mapping that introduced a number of ideas that became widespread in the field, of which we'll give an overview in Section 3.1.5. Later, in Section 4, we also describe the various state of the art TM and ITM methods that we compare our approach with. For comprehensive surveys that include methods not based on vision models we refer the interested reader to [18] and [63].

But despite the fact that there are a number of very effective TM and ITM algorithms, that the field is mature enough, and that the development of new TM methods seems to have peaked a few years back, the cinema industry does not rely on automatic methods and resorts instead to intensive manual process by very skilled technicians and color artists. For accurate on-set monitoring of HDR footage, 3D LUTs are created beforehand by cinematographers and colorists. The first pass of the color grade, called technical grade, where the image appearance is homogenized among shots and made to look as natural as possible [76], is performed manually by the colorist. The mastering and versioning processes are assisted by proprietary automated transforms, the most popular ones coming from Baselight, Resolve, Transkoder/Colorfront and Dolby Vision (which are the ones we will be comparing our framework with), but there is a subsequent and quite substantial input from the colorist, who's required to perform trim passes.

These practices can be explained by the fact that the cinema industry has almost impossibly high image quality standards, beyond the reach of most automated methods, but another, very important point that can be inferred is the following: cinema professionals have the ability to modify images so that their appearance on screen matches what the real-world scene would look like to an observer in it [76]. Remarkably, artists and technicians with this ability are able to achieve what neither state-of-the-art automated methods nor up-to-date vision models can. Put in other words, the manual techniques of cinema professionals seem to have a "built-in" vision model.

We propose in this work effective tone mapping and inverse tone mapping algorithms for production, post-production and exhibition. Our contributions are the following:

- The proposed methods outperform, in terms of visual appearance, the state-of-the-art algorithms from academia as well as the standard industry methods, according to psychophysical validation tests performed by cinema professionals.
- The methods have very low computational complexity and can be executed in real-time.
- We show the limitations of existing objective image quality metrics for TM and ITM, since they do not correlate well with the preferences of observers.
- Our TM and ITM algorithms are based on vision models whose parameters are fine-tuned by cinema professionals. This idea, clearly useful for our cinema applications, might also prove helpful in the opposite direction, for vision science: colorists may assist the improvement or development of novel, more accurate vision models.

In the next section we will start by briefly mentioning a few principles and findings from the vision science literature, some of them quite recent, that are relevant to the HDR imaging problem and that form the basis of our framework.

2 Some vision principles relevant for HDR imaging

One fundamental challenge that our visual system must tackle is imposed by the limited dynamic range of spiking neurons, of around two orders of magnitude, while in the course of a day the ambient light level may vary over 9 orders of magnitude [23]. In order to deal with this the visual system uses a number of strategies.

One of them is adapting sensitivity to the average light intensity. This process, termed light adaptation, is fully accomplished by the retina. It starts already in the photoreceptors, whose sensitivity declines inversely with the ambient light level over a wide range of intensities [68]: the net result is that retinal output can become relatively independent from illumination and encode instead the reflectance of objects, which provides a key survival ability [23].

Another one is encoding local contrast instead of absolute light level. Local contrast is defined as percent change in intensity with respect to the average or background level. In a typical scene large changes in the absolute level of illumination do not have an impact on the range of contrast values, that remain fairly constant and below two orders of magnitude [79], and therefore can be *fit* into the limited dynamic range of neurons.

A third strategy is splitting contrast signals into ON and OFF parallel channels. Visual signals are split at the retina into no less than 12 parallel pathways (also called visual streams or channels) that provide a condensed representation of the scene that can pass through the anatomical bottleneck that is the optic nerve [49]. Some of these channels go all the way from the retina to the primary visual cortex, like the ON pathway, that encodes positive contrast, and the OFF pathway, that encodes negative contrast [83]. That is, ON neurons are mostly responsive to increases in light intensity with respect to the average level, while OFF neurons mainly respond to light intensities that go below the average level. The existence of these two channels has been reported in many animal species, from flies to primates, that use vision to move about in their surroundings [31]. In [27] it's hypothesized that the reason for the existence of the ON and OFF pathways is efficiency: neuron spikes consume energy, so rather than having neurons firing at some significant level for the average light intensity so that the firing rate can decrease for low intensities and increase for high intensities, it appears more convenient to have the neurons be silent or firing at low rates for the average light intensity, with ON neurons devoting their whole dynamic range for increments over the average and OFF neurons for decrements.

These vision properties just described are only three of the many manifestations of efficient representation or efficient coding principles in the visual system, a very popular school of thought within vision science [52] that postulates that all resources in the visual system are optimized for the type of images we encounter in the natural world. Furthermore, this optimization is not static, the visual system must modify its computations and coding strategy as the input stimuli changes, adapting itself to the local characteristics of the stimulus statistics [81].

Adaptation is an essential feature of the neural systems of all species, a change in the input-output relation of the system that is driven by the stimuli [82]. Through adaptation the sensitivity of the visual system is constantly adjusted taking into account multiple aspects of the input stimulus, matching the gain to the local image statistics through processes that aren't fully understood and contribute to make human vision so hard to emulate with devices [15].

There are very different timescales for adaptation [4]: retinal processes that adapt to local mean and variance take place in less than 100msec, the fixation time between consecutive rapid eye movements (microsaccades) [15], while global adaptation processes are usually in the order of a few seconds [23].

In the early visual system, adaptation is concerned mainly with changes in two statistical properties of the light intensity: its mean and variance [14]. Adaptation to the mean is what's referred to as light adaptation. Contrast adaptation tailors the performance of the visual system to the range of fluctuations around the mean, i.e. the contrast, and for instance when the contrast increases (under constant mean) the retinal ganglion cells become less sensitive [4].

By adapting to the statistical distribution of the stimulus, the visual system can encode signals that are less redundant and this in turn produces metabolic savings by having weaker responsiveness after adaptation, since action potentials are metabolically expensive [33]. Atick [3] makes the point that there are two different types of redundancy or inefficiency in an information system like the visual system:

1. **If some neural response levels are used more frequently than others.** For this type of redundancy, the optimal code is the one that performs *histogram equalization*. There is evidence that the retina is carrying out this type of operation [53]: Laughlin showed in 1981 how the photoreceptors of a fly had a response curve that closely matched the cumulative histogram of the average luminance distribution of the fly's environment.
2. **If neural responses at different locations are not independent from one another.** For this type of redundancy the optimal code is the one that performs decorrelation. There is evidence in the retina, the LGN and the visual cortex that receptive fields act as optimal "whitening filters", decorrelating the signal. It should also be pointed out that a more recent work [67] contends that decorrelation is already performed by the rapid eye movements that happen during fixations, and therefore the signal arrives already decorrelated at the retina: the subsequent spatial filtering performed at the retina and downstream must have other purposes, like enhancing image boundaries.

Neural adaptation performs a (constrained) signal equalization by matching the system response to the stimulus mean and variance [15], thus ensuring visual fidelity under a very wide range of lighting conditions.

Figure 1 (left) shows that when the mean light level is high, the nonlinear curve that models retinal response to light intensity is a sigmoid function with less steep slope than when the mean light level is low. Figure 1 (right) shows that at a given ambient level, the slope of the sigmoid is lower when the contrast is higher. The same behavior has been observed for the nonlinear functions that model lightness perception in HDR images [61].

In both cases, the data is consistent with the nonlinearity of the neural response to light performing histogram equalization, since the nonlinearity behaves as the cumulative histogram (which is the classical tool used in image processing to equalize a histogram) does: darker images and images with lower contrast typically have less variance and therefore their cumulative histograms are steeper.

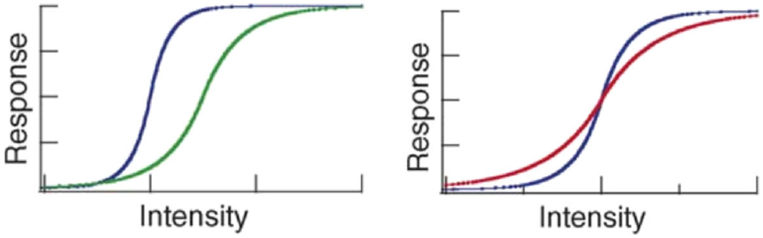


Fig. 1 Neural adaptation to mean and variance. Left: neural response to higher (in green) and lower (in blue) mean luminance. Right: neural response to higher (in red) and lower (in blue) luminance variance. Adapted from [15]

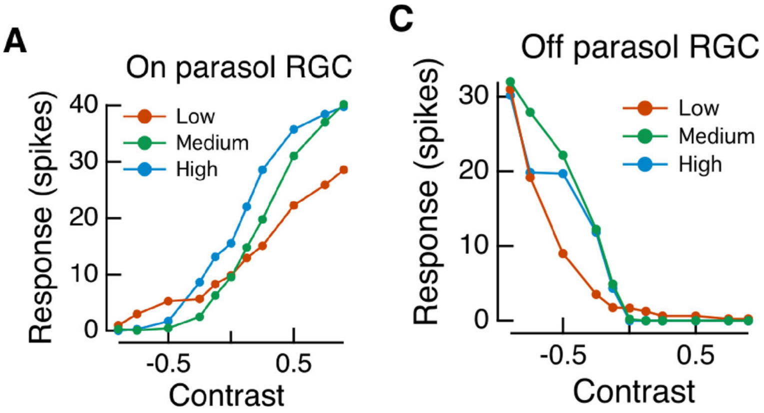


Fig. 2 ON and OFF cells have different nonlinearities. Figure from [73]

We have performed psychophysical experiments where the observers manipulate a display nonlinearity in order to optimize image appearance [32], and our results corroborate that the visual system performs histogram equalization by showing how observers prefer display nonlinearities that allow the displayed image to be perceived as having a brightness distribution as close to uniform (i.e. with an equalized histogram) as possible.

Recent works from neurophysiology prove that OFF cells change their gain more than ON cells during adaptation [55], and, very importantly for our applications, that the nonlinear responses of retinal ON and OFF cells are different [35, 73, 74], see Fig. 2. This data on neural activity is consistent with psychophysical data [84] that demonstrates that our sensitivity to brightness is enhanced at values near the average or background level, so the brightness perception nonlinearity has a high slope at the average level, different arcs at each side of it, and changes shape as the background level changes, see Fig. 3 (left).

Figure 3 (right) shows that, in linear coordinates, the brightness perception curve can't be adequately modeled by a simple power-law function as it's usually done in classic models

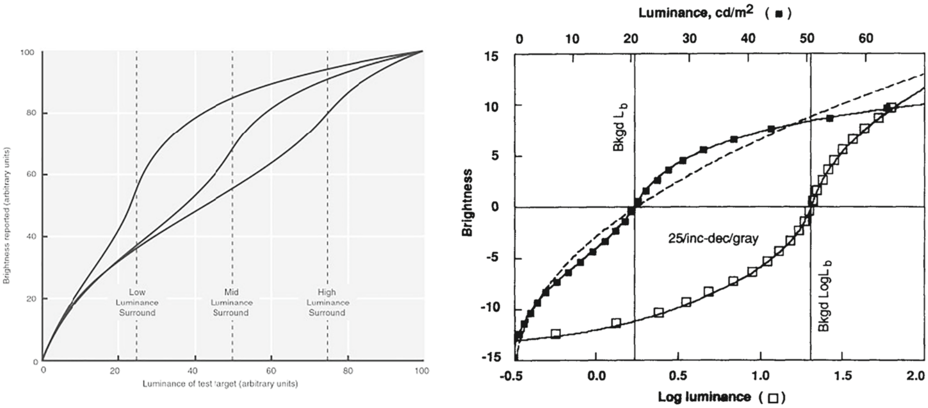


Fig. 3 Left: the shape of the brightness perception nonlinearity is different for values below and above the background level (from [50]). Right: the brightness perception curve is more adequately modeled with two power-laws (linear coordinates) or an asymmetric sigmoid (log coordinates) (from [84]). This psychophysical data is consistent with neural data showing ON and OFF channels having different nonlinearities

(e.g. [71]), and rather it should be represented by two power-laws, one for values below the average level, another for values above it. We can also see in Fig. 3 (right) that when using a logarithmic axis, the brightness perception curve is not a regular, symmetric sigmoid function like the Naka-Rushton equation used to model photoreceptor responses [68].

The consequence of the phenomena just described and depicted in Figs. 2 and 3 is that power-law models of the nonlinear retinal response to the input stimulus should have different exponents for the values below the average light level than for those above the average level.

Once more we have an element that corroborates the efficient coding theory, which is the following. The average luminance histogram of natural images is, in log-log coordinates, piece-wise linear [25], with two lines of different slope at each side of the average level, see Fig. 4 (left). It's easy to see that the cumulative histogram will also be piece-wise linear in log coordinates. Since a linear function in log coordinates corresponds to a power-law in linear coordinates, the retinal nonlinearity modeled as two power-laws can act as the cumulative histogram and therefore it can perform histogram equalization. This histogram equalization process performed by the retinal nonlinearity would be *complete* if the image had a histogram that perfectly matched the piece-wise form of the *average* histogram of natural images, but as this is not the case in general, this nonlinearity performs just an approximation to the full histogram equalization.

The slope or first derivative of the nonlinear retinal response is the gain, and the slope of the brightness perception curve is the sensitivity. They both follow Weber-Fechner's law and decay with increasing stimulus, highlighting again the remarkable similarity between retinal response and perceptual data. Figure 4 (middle) shows that neural gain decreases with light intensity, while Fig. 4 (right) shows that the detection threshold (the inverse of sensitivity) increases with intensity.

Very importantly for our application, sensitivity plateaus for low intensities, suggesting that for the darkest levels there's a limit imposed on the slope of the brightness perception nonlinearity.

At a given ambient level, contrast enhancement is different for achromatic and chromatic signals. A common way of representing vision models is as cascades of linear and nonlinear operations [41], where the linear part is expressed as a convolution with a filter: at retinal level these filters have the form of difference of Gaussians (DoG), so the response to the cell's receptive field surround is subtracted from the response to the receptive field center and this process, termed lateral inhibition or center-surround modulation, produces contrast enhancement. Smaller surrounds enhance more the contrast. Type H2 horizontal cells are the only known retinal cells that sum responses from L, M and S cones [37], their signal is therefore achromatic, and they provide feedback and lateral inhibition to cones. Their surround size is smaller than that of the horizontal feedback to the chromatic

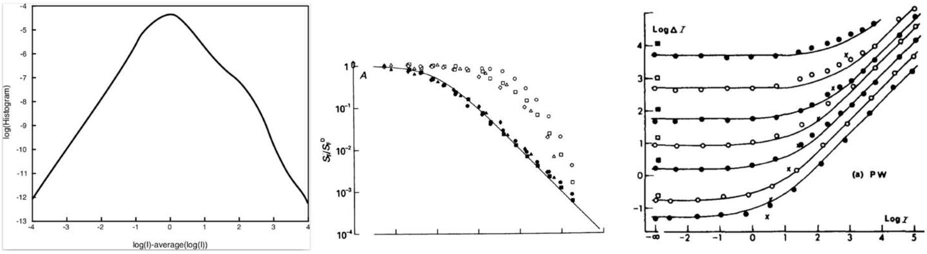


Fig. 4 Left: Average histogram of natural scenes, in log-log coordinates, adapted from [26]. Middle: neural gain vs light intensity (from [44]). Right: JND vs light intensity (from [85])

or color-opponent channels (that encode the difference between cone signals) [69] and as a consequence there is more contrast enhancement for the achromatic channel than for the color-opponent channels.

Coupling among horizontal cells allows them to have, apart from a narrow surround, a much wider surround than was assumed until recently [89], and it's accepted that these cells provide cones with an estimate of average ambient luminance [7] that is not localized but quite global.

Natural image statistics do change with the dynamic range of the scene, but basic retinal transforms (like photoreceptor response and center-surround inhibition at the level of retinal ganglion cells) have been shown to make the signal statistics almost fully independent of the dynamic range [25]. Indeed, most of the vision literature on dynamic range focused on how objects appear invariant over changing lighting conditions, thus omitting irradiance information [45]. But tracking changes in irradiance is advantageous in terms of survival, and there is a type of retinal neuron called ipRGC (for "intrinsically photosensitive retinal ganglion cell") whose outputs are proportional to scene irradiance; individual ipRGCs activate within a range of irradiance values so that collectively they can span and encode light intensity levels from moonlight to full daylight. ipRGCs directly control the pupillary reflex and the circadian rhythm, that modulates the release of dopamine in the retina, which in turn alters the size of the receptive field of horizontal cells performing local contrast enhancement on the cone signals [11]. The key implication is that contrast enhancement will increase as the ambient level is reduced.

As mentioned above, light adaptation takes place fully in the retina, and a substantial part of the contrast adaptation as well. Some of these were phenomena that used to be attributed to cortical processing, but there's now a growing consensus that the retina performs many more tasks and more complicated processes than it was previously assumed [23, 43, 49].

3 Proposed framework

Our framework consists of three methodologies for dealing with HDR material in three different scenarios that cover the cinema production pipeline, from shooting to exhibition. We will start by introducing our approach for TM of ungraded footage (coming directly from the camera) that follows the vision principles enunciated in the previous section. This method can be used for on-set monitoring during the shoot, and for creating the technical grade during post-production. Next we will detail what modifications must be made to this TM algorithm in order to perform TM of graded content. The intended application of this second method is for generating SDR versions from an HDR master. Finally we will introduce our method for ITM, intended to allow HDR cinema projectors and displays to enhance on-the-fly the dynamic range of their SDR input.

3.1 Tone mapping of ungraded footage

The foundations for the TM problem were laid in the seminal works of [72] and [80], that stated how the tone-mapped material presented on a screen should produce "*a subjective experience that corresponds with viewing the real scene*", and therefore the tone-mapping method must behave as "*the concatenation of the real-world observer, the inverse of the display observer, and the inverse of the display model*".

Let H denote the HDR image representing the real-world scene, V_1 a function that represents the real-world observer (in a sense that will be clarified shortly), V_2 a function that represents the display observer, and D a function that represents the nonlinear display response. With this notation, matching the subjective experiences of seeing the real-world scene and the tone-mapped image on the screen is expressed as

$$V_1(H) = V_2(D(TM(H))), \tag{1}$$

from which we get that

$$TM(H) = D^{-1} \circ V_2^{-1} \circ V_1(H) \tag{2}$$

and the basic principle recalled above about how a TM method should behave can be stated as:

$$TM \equiv D^{-1} \circ V_2^{-1} \circ V_1 \tag{3}$$

In our case, we choose to consider V_i, $i \in \{1, 2\}$, as a function that encapsulates the retinal processes of global nonlinear response followed by contrast enhancement that we saw in Section 2 account for light adaptation and contrast adaptation in the retina. Therefore, V_i does not represent the full image percept (it's ignoring higher-level processes that involve for instance edge orientation, memory colors, etc.) and just aims to emulate retinal output. Nonetheless, if two scenarios produce the same retinal output then by necessity they will produce the same subjective experience, so for our application we are not losing generality by defining V_i in this way.

The other choice we make is that our goal is for TM to approximate V_1, and have cinema experts optimize its parameters in a reference environment so that (1) holds: in other words, we propose as tone mapping operator a vision model that is fine-tuned by cinema professionals. From (3), if our goal is met then this implies that $D = V_2^{-1}$, which corresponds to the simplified model used in practice where the display nonlinearity is very similar to the inverse of the brightness perception nonlinearity of the visual system [59].

This similarity between V_2 and D^{-1} is a good approximation that has been established through experiments using simple synthetic stimuli, with no context adaptation and in laboratory conditions. For natural images as it is our case, we conjecture that the adaptation effects, context influence and all other aspects that are required for a perceptual match (1) are handled by the $TM \equiv V_1$ function, *thanks to it having been optimized by cinema professionals.* As we mentioned in the introduction, cinema professionals have the ability to modify images to produce a perceptual match between the real-world scene and the display, and remarkably they are able to achieve what neither state-of-the-art automated methods nor up-to-date vision models can. Put in other words, the manual techniques of cinema professionals seem to have a "built-in" vision model.

In summary, we will propose a TM algorithm that emulates retinal response to a real-world scene. For that we will proceed in three consecutive steps, first emulating light adaptation on a single frame, next emulating contrast adaptation, and finally ensuring temporal coherence on the video sequence.

3.1.1 Light adaptation

The first stage of our TM algorithm takes as input the HDR source and emulates light adaptation by passing the image values through a global response curve. This is a dynamic nonlinearity that adapts to the image statistics, aiming to perform histogram equalization as per the theory of efficient representation, but with some constraints that comply with neuroscience data cited in Section 2:

- The slope is limited for the darkest level.

- For intensity values above the average, the response curve is approximated by a power-law, of some exponent γ^+.
- For intensity values below the average, the response curve is approximated by a power-law, of some exponent γ^-, which is generally different from γ^+.

We estimate the average μ of the light intensity based on the median of the luminance channel Y of the HDR source; the median provides a better fit than the mean to the piece-wise linear shape, in log-log coordinates, of the average histogram of natural scenes [25]. The values for γ^+, γ^- are obtained as the slopes of the piece-wise linear fit to the cumulative histogram of the luminance. In this way they can produce a nonlinearity that, if it were applied to Y, would tend to equalize its histogram, making the distribution of Y closer to being uniform. We define the nonlinearity $\hat{NL}(\cdot)$ to be applied to a normalized signal $C \in [0, 1]$ as a power-law $p(\cdot)$:

$$\hat{NL}(C) = C^{p(C)}, \tag{4}$$

where $p(\cdot)$ smoothly goes from γ^- to γ^+ with a transition at μ of slope n:

$$p(C) = \gamma^+ + (\gamma^- - \gamma^+)\frac{\mu^n}{C^n + \mu^n}, \tag{5}$$

and both n and the dependence of μ on the median have been determined with the help of cinema professionals, as will be discussed in Section 3.4.

For the darkest levels, the curve should not be approximated by a power-law because their derivative tends to infinity as the intensity goes to zero, in the same manner that the gamma correction curve for the Rec.709 standard, that approximates brightness perception, is a power-law for most of the intensity range except for the lower values, where it's a linear function [59]. For this reason we modify $\hat{NL}$ simply by limiting its slope for very low values of C, when $C < \mu/100$. This is key for our application, because otherwise we would be amplifying the noise in dark scenes, a common issue with video TM algorithms [18].

Although the defining parameters μ, γ^-, γ^+ of the resulting nonlinearity $NL(\cdot)$ have been automatically computed from the distribution of Y, the curve NL is not applied on Y but to each color channel $C \in \{R, G, B\}$, thus emulating the achromatic horizontal feedback to cones mentioned in Section 2:

$$C' = NL(C). \tag{6}$$

3.1.2 Contrast adaptation

The second stage of our TM algorithm emulates contrast adaptation by performing contrast enhancement on the intermediate result coming from the light adaptation stage of our method. The motivation is as follows. The real-world scene, source of the HDR material, has an ambient light level that is typically much higher than that of the environment where the screen is located and where the tone-mapped version will be viewed. We saw in Section 2 that when the mean light level is high, the nonlinear curve that models retinal response to light intensity is a sigmoid function with less steep slope than when the mean light level is low. We also mentioned that, at a given ambient level, the slope of the sigmoid is lower when the contrast is higher. From these two elements we conclude that, if we want to emulate at the low ambient level of the screen environment the response to light intensity that is produced at the high ambient level of the real-world scene, we need to increase the contrast of the tone-mapped image, so that when we look at it the retinal response curve becomes shallower, thus emulating the response to the real-world scene. In short, we have to perform contrast enhancement. Furthermore, following the efficient coding principles and the *equalization* role of neural adaptation, the contrast has to be modified so that the resulting image has a variance value closer to that of an image with a uniform distribution.

As the visual system performs contrast enhancement differently on the achromatic signal than on the chromatic channels, we express in a color-opponent space the light-adapted image coming as the output of the first stage of our method. We use for this the *IPT* color-space [16], where *I* corresponds to the achromatic channel and *P* and *T* are the color-opponent channels. Then we perform contrast enhancement by a process that while in essence emulates convolution with a DoG kernel where the extent of the surround is larger for the chromatic channels *P,T* than for channel *I*, thus achieving more contrast enhancement on *I* than on *P* and *T*, in practice is simplified by performing contrast normalization on the *I* channel and leaving untouched the *P* and *T* channels (equivalent to using a DoG with a very wide surround):

$$\begin{aligned}\{I', P', T'\} &= RGB2IPT(\{R', G', B'\}) &(7)\\ I'' &= SI' + \frac{m}{\sigma}(I' - SI') &(8)\\ P'' &= P' &(9)\\ T'' &= T', &(10)\end{aligned}$$

where $RGB2IPT$ is the transform from RGB to IPT color-space, SI' is a smoothed version of I' obtained by convolution with a regularizing kernel (the sum of two Gaussians, which is the dual narrow-wide form of the lateral inhibition produced by retinal interneurons, as explained in [89] and mentioned previously in Section 2; in our case the Gaussians are of standard deviation 5 and 25 pixels), σ is the standard deviation of I', and m is a constant.

3.1.3 Temporal consistency on sequences

In order to tone-map a frame, first we compute temporary values for μ, γ^+, γ^- following the procedure described above; these are averaged with the parameter values from the previous frames, yielding the final parameter values to be applied on the current frame. This is a standard way, in the TM literature, to extend to video processing a TMO devised for still images [8].

In practice the temporal averaging equation we use is the following, but of course many other options are possible:

$$P_f(i) = (30 * P_f(i-1) + P(i))/31, \tag{11}$$

where $P_f(i)$ is the final value of the parameter P (be it μ, γ^+ or γ^-) for frame i after the temporal average of the single-frame estimate $P(i)$ and the previous frame value $P_f(i-1)$.

3.1.4 Computational complexity

The proposed TM method requires computing basic statistics from the luminance histogram in order to define the global nonlinear transform, and performing a convolution with a fixed kernel for the contrast adaptation process. Therefore, the computational complexity of the method is very low. For $2K$ images and with our current, non-fully optimized GPU implementation, the method is taking 90 ms per frame or about 11fps. We are confident that by parallelizing the histogram computation we can bring the implementation to achieve real-time execution at 24 fps.

3.1.5 Related TM methods

A very basic version of our algorithm, with a simpler vision model and no input from cinema professionals, was introduced in [13]. In that work the slope of the nonlinearity is not limited for low intensity values, which produces abundant noise in common scenarios, while the contrast enhancement process is performed in the same way on the *R, G* and *B* channels, which creates problems in the case of highly saturated colors; see Fig. 5 for some examples.

There are some significant novelties in our approach, to be discussed in Section 6, that make our method more effective than the state of the art, as we will show in Section 4.1. Nonetheless, our proposed method is grounded on ideas that are recurrent in the TM literature and in most cases date back to the earliest works of the mid 1990s: vision emulation, retinal models, histogram re-distribution, power-laws, slope limitation, global operations followed by local enhancement, etc.

The following are very brief overviews of the seminal works on vision-based tone mapping that introduced the above-mentioned ideas, all of which have become very extended in the field, e.g. see [63] and references therein:

- We mentioned above how Tumblin and Rushmeier [72] laid the foundations of the TM problem, they posed it as the quest for an operator that causes “a close match between real-world and display brightness sensations”. They based their TMO on approximating brightness perception through a power-law, using the model and perceptual data of Stevens, and modeling the display response with another power-law nonlinearity.
- The work of Ward Larson et al. [80] was also foundational, they defined what they considered to be the two most important criteria for TM and these criteria have been adopted by the community and explicitly or implicitly assumed by all subsequent works:

  1. Reproduction of visibility: an object is seen on the screen if and only if it’s seen in the real scene.
  2. The subjective experience, the overall impression of brightness, contrast and color, is the same when viewing the image on the screen as when viewing the real scene.

  They propose a TM method that performs histogram equalization, but instead of directly using the cumulative histogram as tone curve (which is what regular histogram equalization does) they approximate brightness perception with a logarithmic function, compute the cumulative histogram curve of the brightness image, and then impose constraints on this curve based on models of contrast perception. In this way they aim to comply with the reproduction of visibility requirement, i.e. they don’t want the tone curve to introduce details and contrast that are not visible in the real scene.

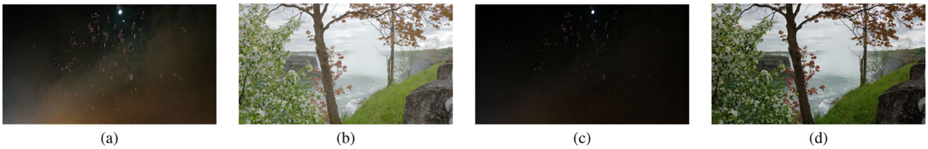


Fig. 5 Images **a** and **b**: example results of a preliminary version of our algorithm [13], with a simple vision model and no input from cinema professionals. Notice the noise and color problems. Images **c** and **d**: results with the method proposed in Section 3.1. Source images from ARRI and Mark Fairchild’s HDR Survey

- Pattanaik et al. [57] used for tone mapping a global nonlinearity that models photoreceptor responses, the Naka-Rushton equation, with parameters that change in time mimicking the visual response during adaptation.
- The TMO by Ashikhmin [2] estimates a local adaptation luminance for each pixel, then local tone mapping is applied based on perceptual discrimination thresholds, and finally there is a contrast enhancement stage.
- The iCAM06 color appearance model of Kuang et al. [36] performs tone mapping with the Naka-Rushton equation in RGB, and converts the result to IPT before applying difference enhancements to the chroma and color channels.

3.2 Tone mapping of graded content

Graded HDR footage has had its contrast and colors modified according to the creator's intentions, and its appearance has been optimized for an HDR display of given peak luminance in a dim or dark surround. The tone-mapping of this source material should preserve the average picture level, not making the images brighter nor darker but simply reducing the intensity of the highlights; the color intent must also be maintained as much as possible.

A TM algorithm must take the above into account, so for TM of graded content we introduce a number of modifications to the method presented in Section 3.1. Let Y_1 be the normalized (i.e. in the range [0, 1]) luminance of the HDR source, μ_1 the median of Y_1, D_1 the peak luminance of the display for which the HDR source material was graded, and D_2 the peak luminance of the reduced dynamic range display where the tone-mapped results will be presented. Therefore, the median luminance (in light intensity units, i.e. without normalization) of the source is $\mu_1 D_1$, and we want the tone-mapped result to have the same median luminance. When normalizing by the corresponding peak luminance, and using a power-law transform of exponent γ, this translates into the following requirement:

$$\mu_1^{\gamma} = \mu_1 \frac{D_1}{D_2},$$

from which γ can be solved as:

$$\gamma = \frac{log(\mu_1) + log(\frac{D_1}{D_2})}{log(\mu_1)}. \tag{12}$$

(In practice, we limit γ to be larger or equal than 0.45: $\gamma = max(0.45, \gamma)$).

So γ is the power-law exponent that must be used for the average light intensity, but as we saw before the exponents are, in general, different for intensities below the average than for intensities above it. We now define γ^+, γ^- in the following way:

$$\gamma^+ = (1-k)\gamma; \quad \gamma^- = (1+k)\gamma, \tag{13}$$

with k a small parameter.

The tone curve characterized by $\mu_1, \gamma^+, \gamma^-$ is applied to Y_1 yielding a tone-mapped luminance image Y_{TM}. Each color channel $C \in \{R, G, B\}$ is tone-mapped resulting in a final output channel C', in this manner:

$$C' = \left(\frac{C}{Y_1}\right)^s * Y_{TM}, \tag{14}$$

where s is a parameter that adjusts the saturation of the output, a common practice in the TM literature.

The contrast adaptation described in Section 3.1.2 for TM of ungraded content is omitted now. The temporal consistency is enforced in the same way as in Section 3.1.3.

3.3 Inverse tone mapping of graded content

Our method for inverse tone mapping of graded content is based on the following ideas:

1. The global curve for inverse tone mapping of graded content should be the inverse of the curve for tone mapping of graded content.
2. The average picture level should be preserved, just increasing the intensity of the highlights.
3. The color intent must also be maintained as much as possible.

From the first item above we see that, since the global curve for TM of graded content is expressed as two power-laws with different exponents for values below and above the average, the global curve for ITM can also be expressed as two power-laws with different exponents for values below and above the average.

From item number 2 above we can find the exponent γ that preserves the median luminance in a way analogous to the one used in Section 3.2, so γ can now be solved as:

$$\gamma = \frac{log(\mu_2) + log(\frac{D_2}{D_1})}{log(\mu_2)}, \tag{15}$$

where μ_2 is the median of the normalized luminance Y_2 of the source SDR graded content, D_2 is the peak luminance of the display for which the source was graded, and D_1 is the peak luminance of the HDR display intended for the output of the ITM algorithm.

We again define γ^+, γ^- as:

$$\gamma^+ = (1+k)\gamma; \quad \gamma^- = (1-k)\gamma, \tag{16}$$

with k a small parameter.

The tone curve characterized by $\mu_2, \gamma^+, \gamma^-$ is applied to Y_2 yielding an inverse-tone-mapped luminance image Y_{ITM}. Each color channel $C \in \{R, G, B\}$ is inverse-tone-mapped resulting in a final output channel C', in this manner:

$$C' = \left(\frac{C}{Y_2}\right)^s * Y_{ITM}, \tag{17}$$

where s is a parameter that adjusts the saturation of the output.

The temporal consistency is enforced in the same way as in Section 3.1.3.

While the literature on inverse tone mapping is not as extensive as that for tone mapping, there are several ITM operators based on inverting sigmoidal compression functions or that take ideas from human vision, e.g. see [5, 28, 65].

3.4 Model tuning by cinema professionals

For the TMO for ungraded content proposed in Section 3.1, input from cinematographers and colorists allowed us to fine-tune the model and to determine an adequate form for the slope n in the nonlinear function NL for the transition between exponent γ^- and γ^+, such that

$$n = -4.5/\mu \tag{18}$$

$$\mu = log(median(Y)) - 2 \tag{19}$$

For the equations in the contrast adaptation section we only need to perform contrast enhancement on the achromatic channel, and an adequate value for m is 0.33.

For the TM and ITM methods for graded content presented in Sections 3.2 and 3.3, we conducted psychophysical experiments in a major post-production facility. Three colorists took part in the tests, where they were asked to match the general appearance of a series of images produced by our methods to the appearance of images produced manually using the current state-of-the-art workflow for TM and ITM of graded content. Images were matched via the adjustment of two algorithm parameters, corresponding to the global contrast and mid-tone luminance value. All the experiments and grading were conducted on calibrated reference displays in a standard reference viewing environment. For the ITM method there was a significant correlation ($R = 0.86$) between the value for γ obtained in (15) and the average value chosen by the colorists, which corroborates the efficacy of the model.

A value of $\gamma^- = 1.4$ provides a good correlation with the observers' settings, from which we can derive k and γ^+ from (16) on an image-by-image basis. Correspondingly, for the TM method for graded content, we set $\gamma^- = 1/1.4$.

The saturation parameter s was set to $s = 0.9$ in (14) and $s = 1.4$ in (17).

Finally, we want to stress that the parameter tuning was performed on images and video sequences that were different from the ones used for validation and comparisons with other methods, to be discussed in the following section.

4 Validation: psychophysical tests

4.1 TM of ungraded footage

We will compare our TMO introduced in Section 3.1 with the three methods that performed best in the very recent survey by [18]:

- Mantiuk et al. [39] proposed a piece-wise linear curve to perform dynamic range compression, where the tone curve parameters are chosen so as to minimize the difference between the estimated response of the human visual system model for the output and the original image. The method can adapt to the particularity of the display and the viewing condition.
- Eilertsen et al. [17] extended the above approach by performing a decomposition into base and detail layers while considering the noise level in the image. The detail layer, after discounting for the noise visibility, is added to the transformed base layer to obtain the final tone mapped image.
- Boitard et al. [8] proposed a two-stage post-processing technique for adapting a global TMO for video. In the first stage, the tone mapped video frames are analyzed to find an anchor frame, then as a second stage the frames are divided into brightness zones and each zone is adjusted separately to preserve the brightness coherence with the anchor frame zone.

We will also compare our TMO with the most popular TM methods used in the cinema industry during production and post-production: camera manufacturer look-up tables (LUTs) (Canon provided LUT, Arri Alexa V3 to Rec709 LUT, and REDRAW default processing in Resolve), Resolve Luminance Mapping, Baselight TM (Truelight CAM DRT).

Some very recent tone mapping works which we would like to compare with in the future are those of Zhang et al. [91], that use a retinal model, and Rana et al. [62], that uses a deep learning approach.

Test content A range of 11 test video clips, of 10 seconds each, have been chosen which possess the following characteristics: natural/on-set lighting, bi-modal and normal luminance distributions, range of average picture level (APL), varying digital cinema capture platforms (i.e. differing capture dynamic ranges), skin tones, memory colors, portraits and landscapes, color/b+w, minimal visual artifacts, realistic scenes for cinema context. Sample frames from the chosen sequences are shown in Figs. 6 and 9. The original test clips were transformed through each of the tone mapping methods listed above.

Observers A total of 17 observers participated in the experiment. This was a diverse group of cinema professionals involved with motion picture production and experienced with image evaluation (digital imaging technicians, directors of photography, colorists, editors and visual effects specialists)

Pairing In order to make an appropriate ranking, every combination of academic methods was paired. Industrial methods were paired directly against our proposed TMO and not with each other. The order of pair presentation was random as to make for a "blind" test.

Observer task and experimental procedure The observer task was pairwise preferential comparison between the outputs of the chosen methods on the grounds of abstract realism (i.e. the "naturalness" of the results). In the experiment, a motion picture professional was ushered into the lab, seated in front of the two monitors, and briefed with the instructions. These explained the cadence and task of the experiment—that observers should use the keyboard controls to select the clip which appears most natural or realistic to them. The instructions also included the stipulation that observers could take as much time as they'd like to view the clips, but that they should view each new clip in its entirety at least once in order to consider all content details in their decision.

To compute accuracy scores from the raw psychophysical data we use the same approach as in [46], that is based on Thurstone's law of comparative judgment, and which we will now describe.

In order to compare n methods with experiments involving N observers, we create a $n \times n$ matrix for each observer where the value of the element at position (i, j) is 1 if method i is chosen over method j. From the matrices for all observers we create a $n \times n$ frequency

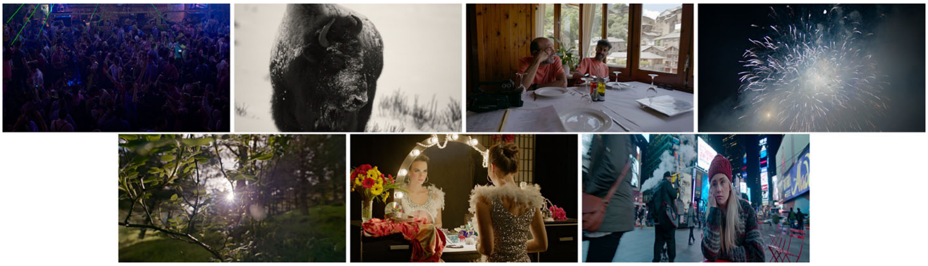


Fig. 6 Sample frames from 7 of the 11 videos used in our validation tests for TM of ungraded content. Samples from the other 4 videos are shown in Fig. 9. Original video sources from [1, 12, 21, 56]

matrix where each of its elements shows how often in a pair one method is preferred over the other. From the frequency matrix we create a $n \times n$ z-score matrix, and an accuracy score A for each method is given by the average of the corresponding column in the z-score matrix. The 95% confidence interval is given by $A \pm 1.96\frac{\sigma}{\sqrt{N}}$, as A is based on a random sample of size N from a normal distribution with standard deviation σ. In practice $\sigma = \frac{1}{\sqrt{2}}$, because the z–score represents the difference between two stimuli on a scale where the unit is $\sigma * \sqrt{2}$ (in Thurstone's paper this set of assumptions is referred to as "Case V"); as the scale of A has units which equal $\sigma * \sqrt{2}$, then we get that $\sigma = \frac{1}{\sqrt{2}}$.

The higher the accuracy score is for a given method, the more it is preferred by observers over the competing methods in the experiment. Figure 7 shows the accuracy scores of the methods from the academic literature, where we can see that our proposed TMO performs best.

For these methods from the academic literature we also performed another experiment where five cinema professionals were asked to rate the outputs on a scale of one to ten in terms of the presence of the following artifacts: noise, haloing, color distortions/clipping and, very importantly for cinema, temporal flickering. A higher rating implies a more noticeable artifact. Observers were allowed to randomly select between a set of video clips produced with different tone mapping methods applied to them, thus eliminating the need for an anchoring system as observers can re-adjust ratings after seeing all the available options. A manually tone mapped version of the clip was also included among the tested methods for observers as a control point. Results are shown in Fig. 8, where we can see that our proposed TMO consistently produces the least artifacts of all the four types, including noise, in which our approach compares well with the method by Eilertsen et al. [17] that is specifically designed to reduce noise in the tone-mapped output.

Regarding the industrial methods, our TMO is preferred over Baselight 66% of the time (with a statistical p-value of $p = 5.4 \cdot 10^{-7}$), over Resolve 73% of the time ($p = 1.4 \cdot 10^{-9}$), and over the camera LUT 58% of the time ($p = 0.0095$).

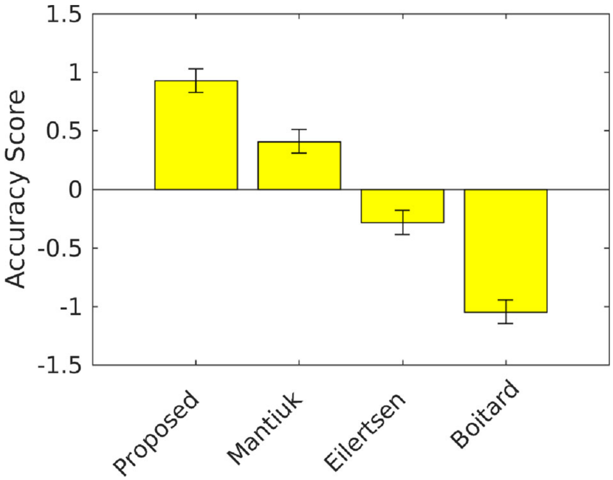
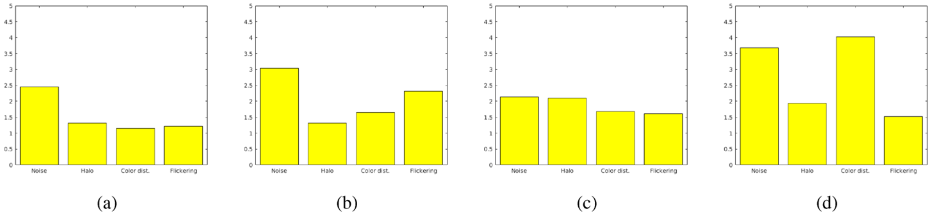


Fig. 7 Observers' preferences of TM methods applied to ungraded content. From left to right: proposed method, Mantiuk et al. [39], Eilertsen et al. [17], Boitard et al. [8]



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

Fig. 8 Observers' rating of artifacts produced by TM methods applied to ungraded content: **a** proposed method, **b** Mantiuk et al. [39], **c** Eilertsen et al. [17], **d** Boitard et al. [8]

Figure 9 shows sample results from the seven TM methods compared on 4 of the 11 test sequences. Notice the improved color and contrast of our results, appearing in the first row.

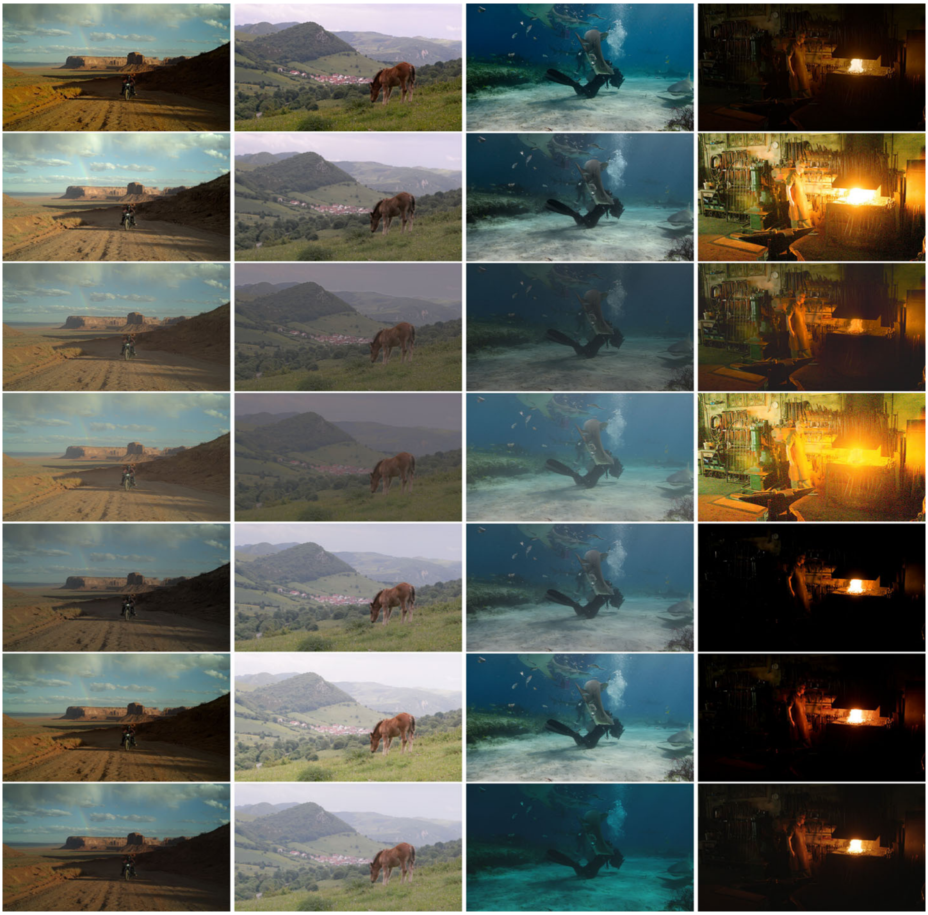


Fig. 9 Sample TM results. From top to bottom: proposed TMO, Mantiuk et al. [39], Eilertsen et al. [17], Boitard et al. [8], Baselight, Resolve, camera LUT. Original video sources from [12, 21, 56]

4.2 TM of graded footage

We will validate the performance of our TM for graded content (TMG) method, introduced in Section 3.2, via pairwise comparison with the methods of [17, 39], our TMO for ungraded content introduced in Section 3.1, and the industrial method of Dolby Vision. Video clips with the various methods applied to them were displayed to observers on a professional SDR reference monitor, who were instructed to choose the one which best matches a manually tone-mapped SDR reference (created by a colorist), in a two-alternative-forced-choice (2AFC) paradigm. This experiment was conducted at Deluxe, the largest post-production house in Spain, where a total of 6 observers, all professional cinema colorists, participated.

Figure 10 shows the accuracy scores of the methods from the academic literature, where we can see that our proposed TMG algorithm performs best (although the inter-subject variability is larger in this experiment than in the one in Section 4.1). It's also interesting to note that our TMO for ungraded content performs worst, highlighting the importance of dealing differently with graded and ungraded content when developing TM methods.

Regarding the industrial method, our TMO is preferred over Dolby Vision 55% of the time; the p-value in this case is $p = 0.25$, not statistically significant, consistent with the very similar performance of both methods.

Figure 11 shows sample results from all the methods compared. Notice how our results, first row, look quite similar to the manual grading results, second row.

4.3 ITM of graded footage

The performance of our method for inverse tone mapping of graded content (ITMG) was validated via pairwise comparison with the Transkoder/Colorfront industry software and three very recent state-of-the-art ITM methods from the academic literature:

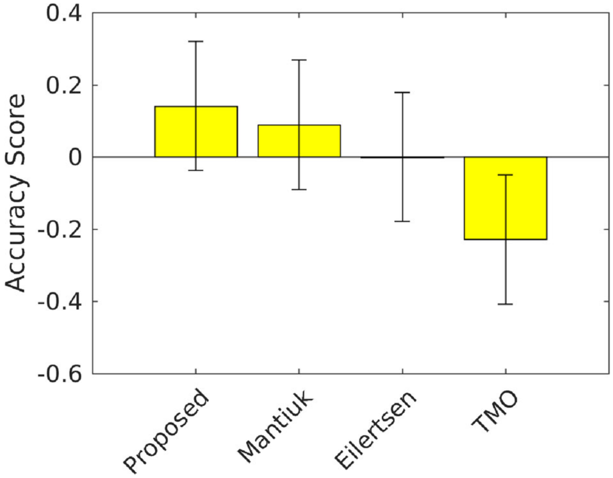


Fig. 10 Observers' preferences of TM methods applied to graded content. From left to right: proposed TM method for graded content, Mantiuk et al. [39], Eilertsen et al. [17], proposed TMO for ungraded content introduced in Section 3.1

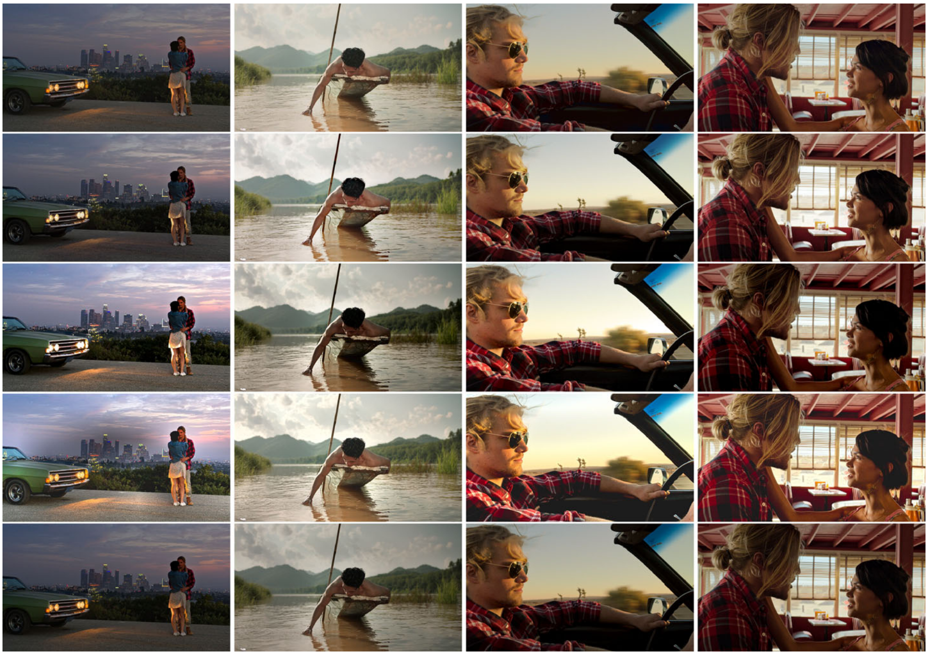


Fig. 11 Result comparison for TM of graded content. From top to bottom: proposed method, manual grading, Mantiuk et al. [39], Eilertsen et al. [17], Dolby. Original video sources from [1]

- Masia et al. [42] proposed a simple gamma curve as the inverse tone mapping operator, where the gamma value is calculated by a linear combination of image key value, the geometric mean of the luminance and the number of overexposed pixels.
- In order to preserve the lighting style aesthetic, Bist et al. [6] also used a gamma curve, which is calculated from the median of the L* component of the input image in the CIE Lab color space.
- Luzardo et al. [38] proposed a simple non-linear function that maps the middle gray value (0.214 for sRGB linear LDR images) of the input image to that of the output HDR image, where the mid-gray of the HDR image is computed by a linear combination of the geometric mean, the contrast and the percentage of overexposed pixels of the luminance channel of the input image.

Let us note that there are also some recent works on ITM with deep learning [19, 20, 30, 86]. They produce results that are not display-referred, meaning that they are not encoded to be reproduced on any particular display and the tone mapping is not designed with any consideration for image appearance in this sense (e.g. the maximum value of the result is not associated to a given peak luminance of a display). Therefore these methods would require an additional tone mapping step in order to present their results on a display and perform a comparison with our method, whose intention as we have made clear is entirely to preserve display-referred appearance between SDR and HDR. Furthermore, one of the main goals of the methods in [19, 20] is to recover image regions where values have been clipped due to over-exposure, but this was not the intention of the film professionals when they created the HDR reference images for our experiments by manual grading of professional, correctly-exposed footage. For these reasons, we decided not to include these methods in our comparisons.

The experiment was conducted at Deluxe with the participation of 7 observers, all of them cinema professionals, and the details are as follows. The reference HDR footage consisted of 10 SDR clips of 10 s each, which were selected and manually graded for a 1,000 cd/m^2 HDR display. Motion picture results from the methods discussed above were displayed to observers on a professional HDR reference monitor, in sequence. Observers were instructed to choose the result which best matches the manually graded HDR reference, shown on a second HDR reference monitor (2AFC).

Figure 12 shows the accuracy scores of the methods from the academic literature, where we can see that our proposed ITM algorithm performs best.

Regarding the industrial method, our TMO is preferred over Transkoder 54% of the time; the p-value in this case is $p = 0.28$, again not statistically significant and consistent with the very similar performance of both methods.

5 Limitations of objective metrics

Evaluating the user preference of TM and ITM results is a challenging problem. Subjective studies like the ones described in the previous sections involve very time-consuming preparations, require special hardware, have to be performed under controlled viewing conditions in an adequate room, etc. Therefore, instead of having to perform user studies, it would be ideal to be able to use image quality metrics to automatically evaluate and rank TM methods, and for this reason a number of objective metrics have been proposed over the years that deal with HDR and tone-mapped images.

In this section we evaluate if any of the most widely-used image metrics for HDR applications is able to predict the results of our psychophysical tests. The metrics we have chosen are TMQI [88], FSITM and FSITM-TMQI [47], HDR-VDP2 [40] and HDR-VQM [48]. FSITM, TMQI and FSITM-TMQI compare the TM results with the original HDR source,

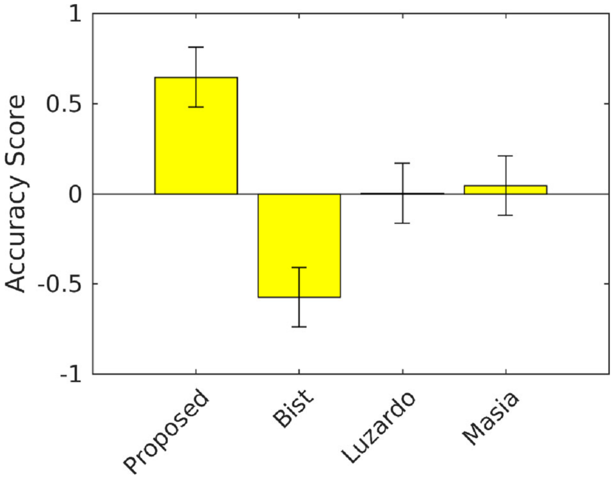


Fig. 12 Observers' preferences of ITM methods applied to graded content. From left to right: proposed ITM, Bist et al. [6], Luzardo et al. [38], Masia et al. [42]

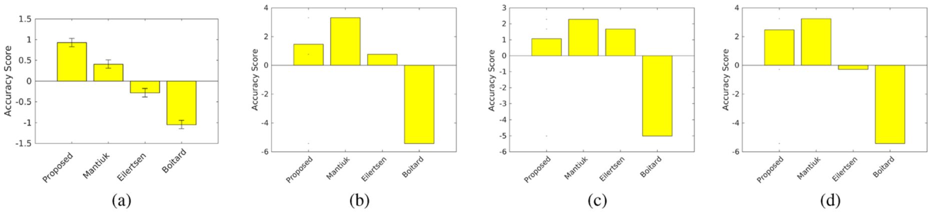


Fig. 13 Validation of metrics for the case of TM of ungraded content: **a** observers' preference, **b** prediction by FSITM, **c** prediction by TMQI, **d** prediction by FSITM-TMQI

while HDR-VDP2 and HDR-VQM compare the manual TM or ITM results with the output of the automatic methods. HDR-VQM was specially developed for video assessment so we have applied it directly on the full test sequences; the other metrics are computed for 1 every 5 frames in each sequence, and intra-frame scores are averaged using TMQI's memory effect.

Furthermore we will also use two very recent, state-of-the-art deep learning metrics for perceived appearance, PieAPP [60] and LPIPS [90]. These two metrics are designed to predict perceptual image error like human observers and are based on large scale data-sets (20K images in one case, 160K images in the other) labeled with pair-comparison preferences and, in the case of LPIPS, using close to 500K human judgments. As these metrics were trained on SDR image data-sets, we'll use them just for the case of TM of graded content.

In order to evaluate the performance of all the above metrics in the context of our experimental setting we will consider the metrics as if they were observers in our pair comparison experiments. This means that, for each metric, we will run all the possible comparisons, and in each comparison we will give a score of 1 to the image with better metric value and a score of 0 to the image with worse metric value. Then, we apply the Thurstone Case V analysis to obtain the accuracy scores that each metric predicts for the TM and ITM methods tested.

Figures 13, 14 and 15 compare the accuracy scores predicted by the metrics with the actual accuracy scores yielded by the observers. These figures show that the metrics were not able to accurately predict the choices made by observers for any application (TM of ungraded or graded content, ITM).

Only in the case of LPIPS (Fig. 14, rightmost plot) it could be argued that the metric is not inconsistent with observers' choices. But Fig. 16 shows examples where LPIPS clearly contradicts observers' preferences.

We are aware that no strong conclusions should be derived from this reduced set of tests, but the limitations of objective metrics that we have observed are in line with the results recently reported in [34].

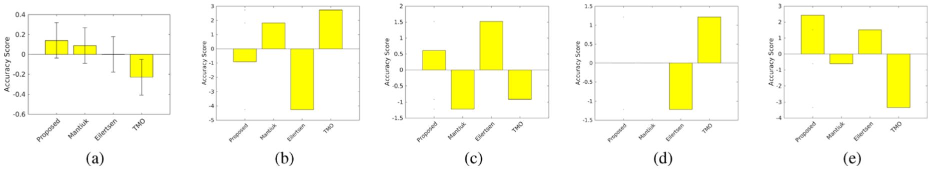
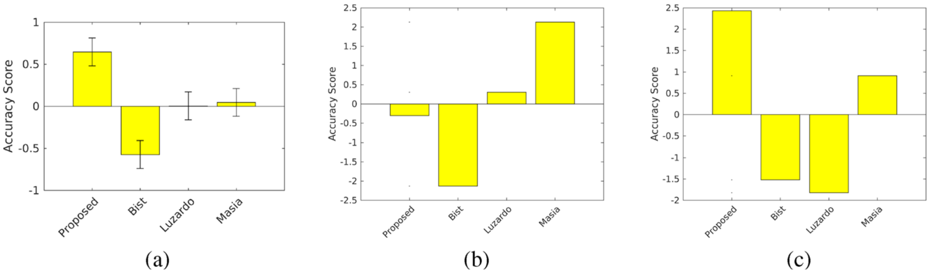


Fig. 14 Validation of metrics for the case of TM of graded content: **a** observers' preference, **b** prediction by HDR-VQM, **c** prediction by HDR-VDP2, **d** prediction by PieAPPs, **e** prediction by LPIPS



(a) (b) (c)

Fig. 15 Validation of metrics for the case of ITM: **a** observers' preference, **b** prediction by HDR-VQM, **c** prediction by HDR-VDP2

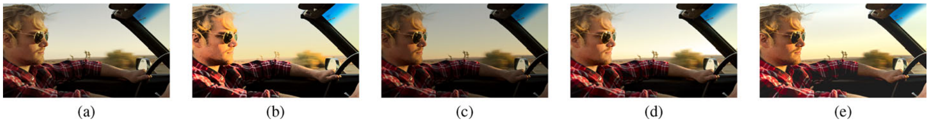
6 Discussion

We would like to share some thoughts on possible reasons for the effectiveness of our approach, as well as its potential for vision science research.

While our proposed TMO for ungraded content is based on a simple retinal model that is consistent with neurophysiological data, it can also be seen as a combination of ideas that have been recurrent in the TM literature, as mentioned in Section 3.1.5. The advantage of our method comes, we believe, from incorporating into our algorithm a number of properties that have been observed in neural response data and that visibly impact the results. Possibly the most important of these properties are:

- The very wide receptive field of horizontal cells supports the use of a global nonlinearity, which is guaranteed to produce no halo artifacts.
- The different response of ON and OFF channels supports modeling the global nonlinearity as an asymmetric sigmoid (in log coordinates) or as two power-laws (in linear coordinates), which allows to better perform histogram equalization.
- The constant neural gain for the lowest light levels supports limiting the slope of the nonlinearity for the darkest values, which allows to reduce the noise.
- The different extent of the horizontal cell feedback for chromatic and achromatic pathways supports performing contrast enhancement differently for the luminance and the chroma channels, which allows for a better reproduction of both color and contrast.

But the core of our method isn't just the accurate modeling of retinal processes: the contribution of cinema professionals like cinematographers and colorists has been essential in order to properly adjust and fine-tune our framework. Colorists are capable of adjusting the appearance of images so that, when we look at them on a screen, our subjective impression matches what we would perceive with the naked eye at the real-world scene. As such, these cinema professionals behave as "human TMOs", complying with the TM



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

Fig. 16 The lower the LPIPS score is, the closest the image is supposed to be to the reference in terms of perception. From left to right: reference (manual grading); output of [8], LPIPS = 0.324; result from Dolby, LPIPS = 0.376; output of [39], LPIPS = 0.39; result from [17], LPIPS = 0.435. Notice how LPIPS considers the second image to be the most similar to the reference, while observers tend to consider the second image as the worst result among these four methods. Original image from [1]

requirements stated in [72, 80]. We have shown in the current work (and we're definitely not the first ones in doing this) how it's appropriate to develop a TMO after a model of retinal processes, and therefore by optimizing our TMO one could argue that the cinema artists have been optimizing a vision model, simple as it might be. It follows that the work of cinema professionals could potentially be useful as well for vision science research, normally limited by the use of synthetic stimuli and restrictive models in a way that has raised important concerns in the vision science community [10, 54]: the input of colorists and cinematographers might help to better find parameter values for the vision model, as they have done in our case for our algorithms, but going forward their contribution could be even more significant, helping to develop and validate models that lie beyond the restrictive setting of L+NL cascades [51, 54].

7 Conclusion

We have proposed a framework for TM and ITM that is useful for the cinema industry over all stages of the production chain, from shooting to exhibition.

Our framework is very simple, its effectiveness stems from adapting data from the vision neuroscience literature to propose a model that is then fine-tuned by cinema professionals that aim for the best perceptual match between on-screen and real-world images.

The algorithms have very low computational complexity and they could be implemented for real-time execution. The visual quality of the results is very high, surpassing the state-of-the-art in the academic literature and outperforming as well (although sometimes by a narrow margin) the most widely used industrial methods. Given that industrial methods are the result of considerable research into very specific tasks using specific technologies and are thus practical solutions to practical problems, that our approach can work as well as these methods across a broad range of problems and technologies indicates that the theoretical underpinning of our methodology has general applicability and operates to a high standard.

The proposed methods have been validated by expert observers, and we have shown that these user preferences can't be predicted by existing objective metrics, highlighting the need to develop more accurate metrics for HDR imaging.

Going forward, we are currently working on adding more features to our basic TM method so that more perceptual phenomena can be simulated.

We are also working on developing a new type of user interface for color-grading software suites that will allow colorists to modify the appearance of images in a way that echoes how the visual system processes information, and in that manner operating the software system becomes (hopefully) more intuitive.

Acknowledgements This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 761544 (project HDR4EU) and under grant agreement number 780470 (project SAUCE), and by the Spanish government and FEDER Fund, grant ref. PGC2018-099651-B-I00 (MCIU/AEI/FEDER, UE). We're very grateful to Albert Pascual, Brett Harrison, Stephane Cattan and everyone at Deluxe-Spain, Alejandro Matus and everyone at Moonlight Barcelona, for their help in fine-tuning and validating our method.

References

1. ARRI (2018) Enhanced capture material for hdr4eu project d2.2. [Online]. Available: https://www.upf.edu/web/hdr4eu/publications
2. Ashikhmin M (2002) A tone mapping algorithm for high contrast images. In: Proceedings of the 13th Eurographics workshop on rendering. Eurographics Association, pp 145–156
3. Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? Netw: Comput Neural Syst 3(2):213–251
4. Baccus SA, Meister M (2002) Fast and slow contrast adaptation in retinal circuitry. Neuron 36(5):909–919
5. Banterle F, Ledda P, Debattista K, Chalmers A (2006) Inverse tone mapping. In: Proceedings of the 4th international conference on computer graphics and interactive techniques in Australasia and Southeast Asia. ACM, pp 349–356
6. Bist C, Cozot R, Madec G, Ducloux X (2017) Tone expansion using lighting style aesthetics. Comput Graph 62:77–86
7. Bloomfield SA, Völgyi B (2009) The diverse functional roles and regulation of neuronal gap junctions in the retina. Nat Rev Neurosci 10(7):495
8. Boitard R, Cozot R, Thoreau D, Bouatouch K (2014) Zonal brightness coherency for video tone mapping. Signal Process: Image Commun 29(2):229–246
9. Boitard R, Smith M, Zink M, Damberg G, Ballestad A (2018) Using high dynamic range home master statistics to predict dynamic range requirement for cinema. In: SMPTE 2018, pp 1–28
10. Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust NC (2005) Do we know what the early visual system does? J Neurosci 25(46):10577–10597
11. Chapot CA, Euler T, Schubert T (2017) How do horizontal cells ‘talk’ to cone photoreceptors? Different levels of complexity at the cone–horizontal cell synapse. J Physiol 595(16):5495–5506
12. Cinema RD (2019) Sample r3d files. [Online]. Available: https://www.red.com/sample-r3d-files
13. Cyriac P, Kane D, Bertalmío M (2016) Optimized tone curve for in-camera image processing. Electron Imaging 2016(13):1–7
14. Demb JB (2008) Functional circuitry of visual adaptation in the retina. J Physiol 586(18):4377–4384
15. Dunn FA, Rieke F (2006) The impact of photoreceptor noise on retinal gain controls. Curr Opin Neurobiol 16(4):363–370
16. Ebner F, Fairchild MD (1998) Development and testing of a color space (ipt) with improved hue uniformity. In: Color and imaging conference, vol 1998, no 1. Society for Imaging Science and Technology, pp 8–13
17. Eilertsen G, Mantiuk RK, Unger J (2015) Real-time noise-aware tone mapping. ACM Trans Graph (TOG) 34(6):198
18. Eilertsen G, Mantiuk RK, Unger J (2017) A comparative review of tone-mapping algorithms for high dynamic range video. In: Computer graphics forum, vol 36, no 2. Wiley Online Library, pp 565–592
19. Endo Y, Kanamori Y, Mitani J (2017) Deep reverse tone mapping. ACM Trans Graph 36(6):177–1
20. Eilertsen G, Kronander J, Denes G, Mantiuk RK, Unger J (2017) Hdr image reconstruction from a single exposure using deep cnns. ACM Trans Graph (TOG) 36(6):178
21. 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, vol 9023. International Society for Optics and Photonics, p 90230X
22. Goeller K (2015) Building the hdr economy nit by nit. In: Annual technical conference and exhibition, SMPTE 2015. SMPTE, pp 1–9
23. Gollisch T, Meister M (2010) Eye smarter than scientists believed: neural computations in circuits of the retina. Neuron 65(2):150–164
24. Gordon A (2018) Beyond better pixels: how hdr perceptually and emotionally effects storytelling. In: SMPTE 2018, pp 1–12
25. Grimaldi A, Kane D, Bertalmío M (2019) Statistics of natural images as a function of dynamic range. J Vis 19(2):13–13
26. Huang J, Mumford D (1999) Statistics of natural images and models. In: IEEE Computer Society conference on computer vision and pattern Recognition 1999, vol 1. IEEE
27. Hubel DH (1995) Eye, brain, and vision. Scientific American Library/Scientific American Books

28. Huo Y, Yang F, Dong L, Brost V (2014) Physiological inverse tone mapping based on retina response. Vis Comput 30(5):507–517
29. ITU-R (2016) Report itu-r bt.2390-0
30. Jang H, Bang K, Jang J, Hwang D (2020) Dynamic range expansion using cumulative histogram learning for high dynamic range image generation. IEEE Access 8:38554–38567
31. Jansen M, Jin J, Li X, Lashgari R, Kremkow J, Bereshpolova Y, Swadlow HA, Zaidi Q, Alonso J-M (2018) Cortical balance between on and off visual responses is modulated by the spatial properties of the visual stimulus. Cereb Cortex 29(1):336–355
32. Kane D, Bertalmío M (2016) System gamma as a function of image-and monitor-dynamic range. J Vis 16(6):4–4
33. Kohn A (2007) Visual adaptation: physiology, mechanisms, and functional benefits. J Neurophysiol 97(5):3155–3164
34. Krasula L, Narwaria M, Fliegel K, Le Callet P (2017) Preference of experience in image tone-mapping: dataset and framework for objective measures comparison. IEEE J Sel Top Signal Process 11(1):64–74
35. Kremkow J, Jin J, Komban SJ, Wang Y, Lashgari R, Li X, Jansen M, Zaidi Q, Alonso J-M (2014) Neuronal nonlinearity explains greater visual spatial resolution for darks than lights. Proc Natl Acad Sci 201310442
36. Kuang J, Johnson GM, Fairchild MD (2007) icam06: a refined image appearance model for hdr image rendering. J Vis Commun Image Represent 18(5):406–414
37. Lee BB, Martin PR, Grünert U (2010) Retinal connectivity and primate vision. Progr Retin Eye Res 29(6):622–639
38. 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: 2018 Picture coding symposium (PCS). IEEE, pp 199–203
39. Mantiuk R, Daly S, Kerofsky L (2008) Display adaptive tone mapping. In: ACM transactions on graphics (TOG), vol 27, no 3. ACM, p 68
40. Mantiuk R, Kim KJ, Rempel AG, Heidrich W (2011) Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In: ACM Transactions on graphics (TOG), vol 30, no 4. ACM, p 40
41. Martinez-Garcia M, Cyriac P, Batard T, Bertalmío M, Malo J (2018) Derivatives and inverse of cascaded linear+ nonlinear neural models. PloS One 13(10):e0201326
42. Masia B, Serrano A, Gutierrez D (2017) Dynamic range expansion based on image statistics. Multimed Tools Appl 76(1):631–648
43. Masland RH (2012) The neuronal organization of the retina. Neuron 76(2):266–280
44. Matthews H, Fain G, Murphy R, Lamb T (1990) Light adaptation in cone photoreceptors of the salamander: a role for cytoplasmic calcium. J Physiol 420(1):447–469
45. Milner ES, Do MTH (2017) A population representation of absolute light intensity in the mammalian retina. Cell 171(4):865–876
46. Morovic J (1998) To develop a universal gamut mapping algorithm. Ph.D. dissertation. University of Derby, UK
47. Nafchi HZ, Shahkolaei A, Moghaddam RF, Cheriet M (2015) Fsitm: a feature similarity index for tone-mapped images. IEEE Signal Process Lett 22(8):1026–1029
48. Narwaria M, Da Silva MP, Le Callet P (2015) Hdr-vqm: an objective quality measure for high dynamic range video. Signal Process: Image Commun 35:46–60
49. Nassi JJ, Callaway EM (2009) Parallel processing strategies of the primate visual system. Nat Rev Neurosci 10(5):360
50. Nundy S, Purves D (2002) A probabilistic explanation of brightness scaling. Proce Natl Acad Sci 99(22):14482–14487
51. Olshausen BA (2013) 20 years of learning about vision: questions answered, questions unanswered, and questions not yet asked. In: 20 years of computational neuroscience. Springer, pp 243–270
52. Olshausen BA, Field DJ (1996) Natural image statistics and efficient coding. Netw: Comput Neural Syst 7(2):333–339
53. Olshausen BA, Field DJ (2000) Vision and the coding of natural images: the human brain may hold the secrets to the best image-compression algorithms. Am Sci 88(3):238–245
54. Olshausen BA, Field DJ (2005) How close are we to understanding v1? Neural Comput 17(8):1665–1699
55. Ozuysal Y, Baccus SA (2012) Linking the computational structure of variance adaptation to biophysical mechanisms. Neuron 73(5):1002–1015
56. Pascual A (2018) Unrealeased footage. [Online]. Available: http://albertpascualcinema.blogspot.com/
57. Pattanaik SN, Tumblin J, Yee H, Greenberg DP (2000) Time-dependent visual adaptation for fast realistic image display. In: Proceedings of the 27th annual conference on computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co, pp 47–54

58. Ploumis S, Boitard R, Jacquemin J, Damberg G, Ballestad A, Nasiopoulos P (2018) Quantitative evaluation and attribute of overall brightness in a high dynamic range world. In: SMPTE 2018, pp 1–16
59. Poynton C (2012) Digital video and HD: algorithms and interfaces. Elsevier
60. Prashnani E, Cai H, Mostofi Y, Sen P (2018) Pieapp: Perceptual image-error assessment through pairwise preference. In: The IEEE conference on computer vision and pattern recognition (CVPR)
61. Radonjić A, Allred SR, Gilchrist AL, Brainard DH (2011) The dynamic range of human lightness perception. Curr Biol 21(22):1931–1936
62. Rana A, Singh P, Valenzise G, Dufaux F, Komodakis N, Smolic A (2020) Deep tone mapping operator for high dynamic range images. IEEE Trans Image Process 29:1285–1298
63. 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, San Francisco
64. Reinhard E, Stauder J, Kerdranvat M (2018) An assessment of reference levels in hdr content. In: SMPTE 2018, pp 1–10
65. Rempel AG, Trentacoste M, Seetzen H, Young HD, Heidrich W, Whitehead L, Ward G (2007) Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. In: ACM transactions on graphics (TOG), vol 26, no 3. ACM, p 39
66. Routhier PH (2018) What are the "killer apps" for hdr? Analysis of sdr assets to predict the potential of hdr. In: SMPTE 2018, pp 1–17
67. Rucci M, Victor JD (2015) The unsteady eye: an information-processing stage, not a bug. Trends Neurosci 38(4):195–206
68. Shapley R, Enroth-Cugell C (1984) Visual adaptation and retinal gain controls. Progr Retin Res 3:263–346
69. Schmidt BP, Neitz M, Neitz J (2014) Neurobiological hypothesis of color appearance and hue perception. JOSA A 31(4):A195–A207
70. SMPTE (2015) SMPTE HDR Study group report. https://www.smpte.org/standards/reports
71. Stevens J, Stevens SS (1963) Brightness function: effects of adaptation. JOSA 53(3):375–385
72. Tumblin J, Rushmeier H (1993) Tone reproduction for realistic images. IEEE Comput Graph Appl 13(6):42–48
73. Turner MH, Rieke F (2016) Synaptic rectification controls nonlinear spatial integration of natural visual inputs. Neuron 90(6):1257–1271
74. Turner HM, Schwartz GW, Rieke F (2018) Receptive field center-surround interactions mediate context-dependent spatial contrast encoding in the retina. bioRxiv p 252148
75. UHDForum (2016) http://ultrahdforum.org/wp-content/uploads/2016/04/Ultra-HD-Forum-Deployment-Guidelines-V1.1-Summer-2016.pdf
76. Van Hurkman A (2013) Color correction handbook: professional techniques for video and cinema. Pearson Education
77. Van Hurkman A (2016) http://vanhurkman.com/wordpress/?p=3548
78. Vandenberg J, Andriani S (2018) A survey on 3d-lut performance in 10-bit and 12-bit hdr bt.2100 pq. In: SMPTE 2018, pp 1–19
79. Wandell BA (1995) Foundations of vision, vol 8. Sinauer Associates, Sunderland
80. Ward G, Rushmeier H, Piatko C (1997) A visibility matching tone reproduction operator for high dynamic range scenes. IEEE Trans Visual Comput Graph 4:291–306
81. Wark B, Lundstrom BN, Fairhall A (2007) Sensory adaptation. Curr Opin Neurobiol 17(4):423–429
82. Wark B, Fairhall A, Rieke F (2009) Timescales of inference in visual adaptation. Neuron 61(5):750–761
83. Wässle H (2004) Parallel processing in the mammalian retina. Nat Rev Neurosci 5(10):747
84. Whittle P (1992) Brightness, discriminability and the "crispening effect". Vis Res 32(8):1493–1507
85. Whittle P, Challands P (1969) The effect of background luminance on the brightness of flashes. Vis Res 9(9):1095–1110
86. Xu Y, Song L, Xie R, Zhang W (2019) Deep video inverse tone mapping. In: 2019 IEEE fifth international conference on multimedia big data (BigMM). IEEE, pp 142–147
87. Yedlin S (2016) On color science. http://www.yedlin.net/OnColorScience/
88. Yeganeh H, Wang Z (2013) Objective quality assessment of tone-mapped images. IEEE Trans Image Process 22(2):657–667
89. Yeonan-Kim J, Bertalmío M (2016) Retinal lateral inhibition provides the biological basis of long-range spatial induction. PloS One 11(12):e0168963
90. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR
91. Zhang X-S, Yang K-F, Zhou J, Li Y-J (2020) Retina inspired tone mapping method for high dynamic range images. Opt Express 28(5):5953–5964