Energy-aware images: Quality of Experience vs Energy Reduction

Olivier Le Meur
InterDigital
Rennes, France
olivier.lemeur@interdigital.com

Claire-Hélène Demarty
InterDigital
Rennes, France
claire-helene.demarty@interdigital.com

Erik Reinhard
InterDigital
Rennes, France
erik.reinhard@interdigital.com

Franck Aumont
InterDigital
Rennes, France
franck.aumont@interdigital.com

Laurent Blondé
InterDigital
Rennes, France
laurent.blonde@interdigital.com

ABSTRACT

The production, transmission and display of video content requires significant amounts of energy. Whether broadcasting or streaming a video, its display on modern televisions is responsible for a significant proportion of the energy consumption. This paper proposes a framework for analysing and processing video frames that allows a modern screen to use less energy when displaying these frames. The content is analyzed prior to encoding and transmission, generating metadata that is attached to the content. After reception by a display, the metadata is then used along with display parameters and user settings to adapt the content prior to display. Two use cases are discerned: one whereby the energy is reduced as much as possible under the constraint that visual quality is maintained, and another where the highest visual quality is sought under the constraint of a fixed reduction of energy use. These use cases would well serve broadcasting and streaming services, respectively.

CCS CONCEPTS
• Computing methodologies → Image processing; Image compression; Machine learning; • Hardware → Power and energy; Displays and imagers.

KEYWORDS
energy, sustainability, attenuation map, JVET

1 INTRODUCTION

Climate change is a fact 1, and every industry should contribute to mitigating its effects by moving to clean sources of energy, and by reducing the amount of energy used. This includes the entertainment and media industry, which serves content to more than 2 billion televisions installed worldwide [3]. Each step, from production, encoding, transmission, to decoding and rendering on television screens contributes to the overall energy consumption, with televisions taking as much as 50% of the total energy use [2, 6]. The reduction of energy consumption of television screens is therefore a priority.

In part driven by regulatory requirements, display manufacturers are implementing various technologies to help reduce the energy consumption of televisions, including for example automatic brightness control. The Green MPEG standard, although regretfully not currently in broad use, offers a metadata solution to globally scale the brightness of a display based on the content received [5], albeit that this does not allow for a flexible pixel-wise content-adaptive approach.

Pixel-wise content-adaptive solutions, however, would offer better opportunities to deliver energy-efficient content to televisions. Most existing pixel-wise content-adaptive technologies aim for a suitable trade-off between the quality of experience (QoE) and energy reduction [7]. This could be achieved for example by linear scaling followed by a (deep-learning-based) contrast enhancement step [9, 12].

The purpose of the present paper is to discuss a framework by which pixel-wise content-adaptive techniques could be incorporated into broadcast and streaming solutions, and to provide an initial comparison between a simple linear scaling and a state-of-the-art pixel-wise content-adaptive method.

The paper is organized as follows. In Section 2, we present the general framework that could be used to reduce the energy consumption of display devices. Two use-cases are presented as well as recent activities in JVET standardization. Section 3 presents and evaluates one recent technology that allows to produce energy-aware images. Section 4 draws some conclusions and presents future work.

2 PIXEL-WISE CONTENT-ADAPTIVE ENERGY REDUCTION

2.1 General framework

For backlit display technologies, the energy consumption correlates strongly with the level of backlight illumination, whereas for emissive displays such as OLEDs, it relates to the pixels’ individual amount of emitted light. In either case, reducing the luminance levels of the content allows a reduction of energy use of display devices. However, for backlit displays, this reduction would go through a backlight scaling algorithm, while for OLED displays, it can be envisioned as a direct and pixel-wise reduction of each pixel’s consumption.

In the latter case, a pixel-wise content-adaptive strategy for energy reduction would exploit the inability of humans to perceive certain changes that can be made to imagery, based on the content itself. Such an approach necessarily consists of an analysis step whereby the content is analysed for its ability to be reduced in luminance while minimizing the loss of visual quality. In a second step, the content would be adapted accordingly, and would also take into consideration the target display device. As the analysis step may be expensive, a pixel-wise content-adaptive energy reduction framework would place this step pre-encoding, while the content would be adapted after reception by a display device. This means that the results of the analysis step need to be communicated to the display device, in the form of metadata. An illustration of such a framework in the global video chain is proposed in Figure 1. Pixel-wise energy-aware image algorithms generating pixel-based information of energy reduction would therefore be mainly positioned prior to the encoding in the analysis step. They would also benefit from several well-suited characteristics: by nature, they are more flexible as changes applied on content can be spatially localized; perceptual properties of the human vision can be exploited to control the content modification at the pixel level; this will allow for a finer management of the trade-off between energy reduction and quality of experience in the resulting modified images. An implementation of this framework was recently proposed as a contribution in the video standards, and is described in the following section.

2.2 Implementation in standards

The standard ISO / IEC 23001-11 Energy-Efficient Media Consumption (Green Metadata) [5] already specifies metadata either to facilitate the reduction of energy usage during media consumption or to reduce power consumption at the display side through display adaptation to content.

As far as the use case of display adaptation is concerned, the corresponding metadata are particularly well tailored to transmissive display technologies with backlight illumination such as LCD displays. These metadata are designed to obtain display energy reduction through e.g., the adaptation of the backlight to specific content characteristics. To this end, they comprise RGB-component statistics and quality indicators of the transmitted video content. Such statistics and metrics can be used to set display controls, and specifically the backlight level, for example through RGB picture components global rescaling. Their use should enable to set the best compromise between backlight/voltage reduction and picture quality or to ensure that either the desired quality level or the display power reduction are reached.

If these metadata already convey some information linked to the reduction of energy consumed by displays, they have the drawback of being tailored to backlit display technologies and therefore only convey global statistics derived from the input content. If it is assumed that the power consumption of OLED screens is linear with luminance, such global statistics can also be helpful in the context of OLED screens. However, this global information is far from optimal and turns out to be limited when applied to more controllable directly emissive displays. A more precise control both on the energy reduction and the quality of experience is indeed possible by modifying locally the content.

To enable an effective implementation of the content-adaptive and pixel-wise framework proposed in the previous section, a recent JVET standardization contribution was proposed [10]. It differs with the Green Metadata one in the sense that it proposes the signaling of new metadata for enabling the attenuation of each sample value (of each color component) of a decoded video at the receiver side. To this end, this contribution proposes the transmission of an SEI message as well as an attenuation map, defined as a new type of auxiliary picture. The attenuation map could be the output of the pre-encoding analysis step (as referred to in Figure 1). It will provide a pixel-wise information of the modification to apply to any pixel of the corresponding input image. The accompanying SEI message will help define how the attenuation map should be applied on the image to be displayed, the simplest way being a pixel-wise subtraction of the attenuation map. The SEI message also conveys other useful information, i.e., on which components (luminance, R, G, B, etc.) the attenuation map is to be applied, indication of any rescaling algorithm to apply to the attenuation map before usage, etc. Together, these enable an intelligent spatially varying reduction of pixel values so as to reduce the energy consumption, while maintaining or controlling the quality of experience for the end user. Furthermore, it is also possible to convey information through this SEI message on how to use the attenuation map to derive a backlit scaling that could serve to reduce the energy consumption of transmissive displays.

This contribution to JVET standard is a direct implementation of the global framework proposed in section 2.1, which enables the transmission of information between the analysis and adaptation steps. We further discuss in the following two different use-cases potentially using this global framework.

2.3 Strategies for building energy-aware images

We see two complementary strategies for implementing content-adaptive energy reduction methods. The first one is conservative, in that it would predict the maximum luminance reduction possible for a given content under the constraint that the visual quality is not affected by more than a given threshold. Such an approach could be used for critical viewing applications, such as high-end movie broadcast. Its implementation in the above framework, with an analysis step prior to encoding, would allow the creators’ artistic intent to be respected as the building of the attenuation map could be done during the content creation, under their direct supervision.
We present in this paper recent methods that aim to compute $\gamma$. An interesting point to underline is that such methods do not require any annotated ground truth. The only required ingredient is the dimming map. As illustrated by Figure 2, the network computes a dimming map which is added back to the luminance channel. The backbone of the proposed network is similar to the one proposed in [12]. It consists of multiple consecutive layers. In order to reduce the computational load, a $2 \times 2$ average pooling is performed after a first convolutional layer with 32 feature maps and kernels of $3 \times 3$ size. A multiscale CAN (Context Aggregation Network) [14] with 32 feature maps and kernels $3 \times 3$ is then used; the dilation rate increases exponentially with the layer number. The output image is obtained through a convolutional layer with 32 feature maps and $3 \times 3$ kernels, followed by linear transformation ($1 \times 1$ convolutional with no non-linearity). The leaky rectified linear unit (LReLU) is used as an activation function in all convolutional layers. Note that the saving rate $R$ is added to the first layers to condition the network on the targeted energy reduction rate. In [9], three loss functions are used, a power loss to maintain the energy reduction, a SSIM-based loss to ensure a good quality and a contrast loss to locally improve the image contrast. We designed our own implementation of [9] which requires 1.8 million of trainable parameters. We retrained from scratch the network with our own strategy as described in the following section.

3.2 Training protocol

To train the network, we linearly combine four loss functions: the MAE loss $L_{MAE}$, the SSIM loss $L_{SSIM}$, the power loss $L_{pow}$ and the Total Variation (TV) loss $L_{TV}$ functions. These losses are described below, where $Y$, $\hat{Y}$, $P_Y$, $P_{\hat{Y}}$ and $DM$ represent the input and output image luminances, the power for the input and output images and the dimming map, respectively.

- $L_{MAE} = \frac{1}{N} \sum_{i=1}^{N} |Y_i - \hat{Y}_i|$, where $N$ is the total number of pixels and $i$ is the spatial coordinates of the pixel.
- $L_{SSIM} = 1 - SSIM(Y, \hat{Y})$.
- $L_{pow} = ||P_{\hat{Y}} - (1-R) \times P_Y||^2$, with $R \in [0,1]$ the target energy reduction rate and $R \times P_Y$ the desired target power.
- $L_{TV} = \frac{1}{N} \sum_{i=1}^{N} (\nabla_{v} DM_i)^2 + (\nabla_{h} DM_i)^2$, $\nabla_{v}$ and $\nabla_{h}$ represent the vertical and horizontal gradients. This loss aims to get smooth dimming maps, easier to transmit as metadata.

R-ACE network is trained on the BSD dataset [8]. This dataset is composed of 300 images, 200 training images, 40 validation images and 60 for testing. Images have a resolution of 481 × 321, in a landscape or portrait format. We randomly crop images into 128 × 128 pixels and 60 for testing. Images have a resolution of 481 × 321, in a landscape or portrait format. We randomly crop images into 128 × 128 pixels.
patches which undergo random data augmentation (i.e., horizontal flip, vertical flip and rotation of 90 degrees). The network is trained using the following parameters: ADAM solver, learning rate of $1 \times 10^{-3}$, weight decay of $1 \times 10^{-5}$, batch size of 4.

During the first 2 epochs, to ensure the QoE of the output image, the loss function is only composed of the two first losses $L_{MAE}$ and $L_{SSIM}$. The training phase converges quickly with a very good quality of reconstruction; the average PSNR value is above 50 dB. After these first epochs, we add the $L_{pow}$ and $L_{TV}$ losses, to further ensure the energy reduction and the smoothness constraint on the dimming map. The coefficients of the linear combination were empirically set to $\{\alpha_{MAE}, \alpha_{SSIM}, \alpha_{TV}, \alpha_{pow}\} = \{0.5, 1.0, 10, 2000\}$.

### 3.3 Performance

#### 3.3.1 On still images. Figure 3 presents the performance of R-ACE network over 60 test images of the BSDS300 dataset. The first 3 graphics present the PSNR, SSIM [13] and LPIPS [15] scores distribution against energy reduction rates. We further measured with a power meter the energy consumption of the rendering of the processed images on a SONY OLED KD - 55AF9 display (55", HD) screen. The last graphic of Figure 3 presents the distribution of energy consumption of tested images as a function of the target energy reduction. The left-hand side plot shows the estimated power consumption of processed images (the estimated power consumption is computed thanks to the linear energy model) whereas the right-hand side gives the actually measured energy consumption of tested images when displayed onscreen.

From Figure 3, we can draw some observations. First, the objective quality is decreasing when the energy consumption target is increasing. For 5% to 10%, PSNR and SSIM are very high (almost 1 for the SSIM). When the rate is higher than 10%, the objective quality decreases. As expected, the PSNR, which is a metric qualifying reconstruction quality at signal level, is much more sensitive to the dimming process than the perception-based SSIM metric; for $R=40\%$, the average PSNR is around 20dB which indicates a poor quality whereas the average SSIM score remains high (around 0.96). Similarly, from the LPIPS average scores, it is observed that the similarity between the original and modified images remains high for for 5% to 10%, and decreases for higher values.

Concerning the energy model, we observe on the right-hand side of Figure 3 that there is a significant difference between the energy reduction target we want to reach and the energy reduction we actually measure when images are displayed on screen. One explanation comes from the simplicity of the energy model we use (see equation 1). Indeed, the linear assumption between energy consumption and luminance level holds for RGB OLED TV screens only. Our OLED screen is an RGBW OLED screen for which such an assumption does not hold. Modeling more precisely the power consumption of displays would be an important axis of research to improve the performances of future energy aware images methods.
Table 1: VMAF and PSNR scores for different video sequences and reduction rates. We report the scores for the R-ACE and linear scaling methods, as [R-ACE, Linear Scaling].

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Features</th>
<th>VMAF R=10%</th>
<th>VMAF R=20%</th>
<th>VMAF R=40%</th>
<th>PSNR R=10%</th>
<th>PSNR R=20%</th>
<th>PSNR R=40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBC_ThemePark_Part1</td>
<td>1080p50, 420, #600</td>
<td>[94.1; 90.7]</td>
<td>[94.1; 78.1]</td>
<td>[62.1; 51.8]</td>
<td>[37.03; 35.99]</td>
<td>[29.24; 30.23]</td>
<td>[24.54; 23.87]</td>
</tr>
<tr>
<td>BBC_ThemePark_Part2</td>
<td>1080p50, 420, #369</td>
<td>[95.54; 99.12]</td>
<td>[99.72; 94.12]</td>
<td>[74.19; 69.49]</td>
<td>[30.60; 30.58]</td>
<td>[25.62; 24.57]</td>
<td>[18.06; 18.08]</td>
</tr>
<tr>
<td>BBC_ThemePark_Part3</td>
<td>1080p50, 420, #489</td>
<td>[96.96; 93.65]</td>
<td>[95.64; 81.07]</td>
<td>[64.29; 54.27]</td>
<td>[37.57; 36.43]</td>
<td>[29.72; 30.68]</td>
<td>[25.56; 24.32]</td>
</tr>
<tr>
<td>EBU_Aloha</td>
<td>1080p50, 420, #500</td>
<td>[99.97; 99.65]</td>
<td>[99.97; 87.68]</td>
<td>[72.59; 60.89]</td>
<td>[34.16; 34.00]</td>
<td>[27.95; 28.12]</td>
<td>[21.75; 21.68]</td>
</tr>
<tr>
<td>EBU_dance</td>
<td>1080p50, 420, #502</td>
<td>[98.84; 97.77]</td>
<td>[98.13; 86.06]</td>
<td>[61.26; 58.61]</td>
<td>[32.20; 32.32]</td>
<td>[26.87; 26.36]</td>
<td>[19.67; 19.89]</td>
</tr>
<tr>
<td>NTT_BQTerrace</td>
<td>1080p60, 420, #601</td>
<td>[96.9; 93.5]</td>
<td>[98.51; 81.11]</td>
<td>[68.4; 54.3]</td>
<td>[31.04; 31.07]</td>
<td>[25.97; 25.06]</td>
<td>[18.51; 18.57]</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>[97.6; 91.78]</td>
<td>[97.56; 84.68]</td>
<td>[67.13; 58.22]</td>
<td>[33.76; 33.39]</td>
<td>[27.56; 27.50]</td>
<td>[21.33; 21.06]</td>
</tr>
</tbody>
</table>

Figure 4 presents some illustrations of energy-aware images produced by the implemented method. For the sake of completeness, we also present the attenuation map for different energy reduction rates. When the energy reduction rate increases, we observe that images are getting darker, and consequently less energy demanding. The last row illustrates the attenuation maps obtained for R=40%.

3.3.2 On video sequences. VMAF [11] is a full-reference perceptual video quality assessment algorithm developed by Netflix. The VMAF score is computed between some original video sequences from the JVET test sequences [1] and their energy aware processed correspondences. For the sake of comparison, VMAF scores are also provided for processed sequences on which a linear scaling of the luminance was applied. It simply consists in determining a scaling coefficient $k$ to reach an energy consumption target $R$, such that:

$$R = 1 - \frac{\hat{P}}{P}$$

(2)

where $\hat{P}$ and $P$ represent the powers dissipated by our OLED screen when displaying the processed and the original sequences, respectively. Assuming that the energy model is linear, the scaling coefficient $k$ is given by:

$$k = (1 - R)^{1/\gamma}$$

(3)

Where, $\gamma$ is the gamma correction of the screen.

Table 1 reports the VMAF scores. As expected, they decrease with the target energy reduction rate, going from a score higher than 90 for $R=10\%$ to a score of 60 for $R=40\%$. The R-ACE network also outperforms a simple linear scaling and the difference between the two methods increases when $R$ gets higher.

4 CONCLUSION AND DISCUSSION

In this paper, we present a framework aiming to reduce the energy consumption at the display side. The proposed framework can be incorporated into broadcast and streaming solutions. The main idea is to compute an attenuation map at the encoder side and to transmit it along the video chain. Specific SEI message is proposed to define how to transmit and to use such map. The display device can then use such attenuation map in order to reduce the amount of emitted light while maintaining or controlling the QoE.

To illustrate this proposal, we use an existing deep-learning model to populate the proposed SEI message in a context of video streaming. From an input image, it computes an attenuation map for different energy reduction rates. We then present an objective assessment and a comparison to the simple linear scaling approach. Results are promising especially when the energy reduction rate is increasing. There are still however several obstacles to overcome. First, the energy model simulating the power consumption of the display is not accurate enough. We observe a discrepancy between the target rate of energy reduction and the one we achieve. Second, in order to reduce the complexity at the display side, we have to transmit side-information all along the video chain. Although they were proposed in a first JVET contribution, they would require to be standardized. These two points will be addressed in future work.

5 ACKNOWLEDGMENTS

This work has been achieved in the context of the project 3EMS-2 funded by the “Région Bretagne”, Rennes Métropole, co-funded by E.U and supported by “Images et Réseaux”.

REFERENCES


Figure 4: From top to bottom: first row, original images from BSDS dataset [8]; the second to fourth rows correspond to an energy reduction rate of 10%, 20% and 40% respectively. Last row: attenuation maps for $R = 40\%$.


