Subjective quality of videos displayed with local backlight dimming at different peak white and ambient light levels

Claire Mantel, Jari Korhonen and Søren Forchhammer
Department of Photonics Engineering
Technical University of Denmark
2800 Kongens Lyngby, Denmark

Jesper Pedersen +
+ Bang & Olufsen A/S,
Peter Bangs Vej 15,
7600 Struer, Denmark

Søren Bech ++, *
* Section of Signal and Information Processing
Department of Electronic systems,
Aalborg University
9100 Aalborg, Denmark

Abstract—In this paper the influence of ambient light and peak white (maximum brightness) of a display on the subjective quality of videos shown with local backlight dimming is examined. A subjective experiment investigating those factors is set-up using high contrast test sequences. The results are firstly analyzed with an ANalysis Of VAriance, which shows significant effect of the Sequence and Algorithms interactions and the Peak White and Ambient Light interactions. This latter allows us to define test cases for which the peak white and ambient light are adapted to one another. An objective model of the subjective grades is then computed using Partial Least Squares regression that achieves 0.68 correlation.

I. INTRODUCTION

Historically much of research on quality assessment has focused on compression and transmission artifacts [1] as they are the most common defects in an image processing chain. Clearly, other aspects are to be considered for a more comprehensive view on quality assessment, as is defined by the term Quality of Experience by the Qualinet group [2]. Obviously the relative level of each type of defect determines its impact on the overall quality.

The focus of this paper is high quality videos (meaning that no compression or transmission artifacts are considered). In this framework other factors become predominant, e.g. the rendering of the display. The two factors, part of the viewing conditions, investigated here are the maximum brightness, or peak white, of the display and the ambient light. In most cases for quality assessment the display is considered transparent regarding the quality and recommendations for subjective quality experiments only advise minimum characteristics of displays [3], [4]. Nonetheless conventional LCDs do not render frames in perfect quality. One example is the leakage defect: as the Liquid Crystal (LC) cells are not able to completely block the light coming from the backlight, the minimum black level of an LCD is usually comprised between 0.1-1Cd/m². These levels are perceptibly higher than 0 under not too bright viewing conditions and are responsible for the grayish appearance of black on conventional LCDs.

In order to take the rendering of the display into account for quality assessment a display model is necessary. It can be used as a preemptive step for applying of quality metrics [5], [6].

Local backlight dimming is a technology that aims at saving energy and improving quality of the rendered image at the same time. It consists of decreasing the intensity of backlight segments in areas where a frame is dark, thus saving power and reducing the leakage. The potential drawback of local backlight dimming, called clipping, is not providing enough light to a pixel for it to reach the desired luminance.

One possibility to broaden the scope of quality assessment is to turn to the environment surrounding the observer assessing the quality of a video. Considering the display is a first step, but many other factors play a role: e.g. the characteristics of the room (wall color, type of seat, etc), the distance to the screen, the location of the observer in respect to the display (angle of vision), etc.

This study investigates the impact of the ambient light and the peak white (i.e. the maximum intensity) of the display, as part of the viewing conditions, on the perceived quality. It is known that those two factors influence the abilities of the Human Visual System [7], [8] but their impact on the quality is less known. As the complete light received by the eye of an observer watching a display is determined by not only what is shown on the screen (content and peak white) but also by the light surrounding the display, those two factors are intimately related. Indeed, high-end television models often contain an automatic mechanism adapting the peak white of the display to the ambient light.

In this paper we examine the quality of high contrast videos displayed on a screen with local dimming capability at various ambient light and peak white levels. The aim is here to evaluate if and how the subjective quality is impacted by those factors and propose a new model to account for them. The remaining of the paper is organized as follow: first the relevant literature is presented in Sec. II, then the subjective experiment set-up is detailed in Sec. III, and finally the results are analyzed and a model is proposed in Sec. IV, before the concluding remarks in Sec. V.
The term ambient light usually encompasses both the light from the room reflected on the screen and the light included in the field of view besides that of the screen (called surrounding light) [4], [9]. Although it is possible to model the ambient light in a perceptually accurate way [10], characterizing it [11] and computing it is still quite complex and subjective testing are mostly set up in actual light laboratories (in contrast to modeled environments).

There are different ways to approach the investigation of ambient light and peak white for viewing videos on displays: through its influence on the performance for a given task [12], [13], [14], the visual fatigue [7] or the observer preference [8], [15].

The task can consist in low level action such as contrast threshold perception [14] or higher level applications such as recognition of characters [13] or medical abnormalities [16]. In all cases, the brighter the ambient light, the lower the contrast perceived and the lower the performance. Therefore the ambient light highly influences the outcome of a task. It is for example the case for radiology applications where the viewing conditions can vary consequently from one examination room to another or over time [17] and the visible contrast highly influences the performance [13]. In their study on the influence of ambient light on fracture detection on radiology images, McEntee et al. [12]) show that with lower ambient light (25 or 40 lux) observers perform significantly better than with lighting at 100 or 480 lux.

In [16], peak white is found to have no effect on the detection performance at the three ambient light levels tested (0, 50 and 460 lux) and the authors found no interaction with ambient light level.

However quality has a different scope than performance: indeed in the case of quality concepts such as comfort or fatigue [8], [16] become relevant. The distance of viewing and recommended ambient light level is also different in the two cases: World Health Organization advises ≤100 lux for 30 cm for viewing radiology images whereas ITU recommends 15% of the display peak white and 3 times the height of the display for quality assessment experiments. In [16], the authors show that the self-evaluated fatigue increases significantly with both the ambient light and the peak white. Even though those results are not directly applicable to quality assessment as the task is different, fatigue also plays a role for quality assessment.

To the best of our knowledge there is currently no published quality metric that accounts for the ambient light or the display peak white, to the exception of the HDR-VDP-2 [18] that can contain a display model and can take the luminance of each pixel of the displayed image as input.

Partial Least Squares Regression is a statistical analysis method that allows building a model of a response variable based on a set of numerous and potentially correlated features (the predictors). It was previously used in the context of video quality assessment to build objective metrics by [19] to predict the quality of h.264 encoded videos using objective features extracted from the bitstream and by [20] to predict the quality of videos rendered with various backlight dimming algorithms. It is particularly adapted for the exploratory aspect of this study as the relationship between quality and peak white and ambient light is largely undetermined.

**III. DESCRIPTION OF THE SUBJECTIVE EXPERIMENT**

We set up a subjective experiment to investigate relations influencing contrast display and perception. The independent variables are the sequences (5 elements), the backlight dimming algorithms (2 elements), the peak white of the display (3 levels) and the ambient light (3 levels). The measured variable is the subjective quality of the displayed videos.

Five sequences all containing dark areas that could show leakage on the LED backlit LCD were used. The selected sequences (duration 5-8 s) are: ‘Stars’ [21], ‘Titles’ [21], ‘Volcano’ [22], ‘TchDwn’ [22] and ‘Uboat’; frames of the first four are depicted in Fig. 1, whereas the fifth cannot be shown due to copyright reasons. All sequences have Full HD (1920x1080) resolution and are available in original or extremely good quality (i.e. blu-ray).

![Fig. 1: A frame extracted from four of the five sequences used in the experiment](image1)

Two different backlight types were applied for each sequence: the backlight at full intensity (FBL, corresponding to a conventional LCD) and the block based gradient descent (GD) [23], which is the optimal backlight in terms of Mean Squared Error (MSE) as evaluated by a display model and was found to have high subjective contrast in a previous experiment [24]. As varying the intensity of a LED can produce a flashing defect called flicker, the GD algorithm includes a flicker control mechanism to ensure that no such defect appears by keeping the LED variations below a chosen threshold. The GD algorithm varies the local backlight intensity both spatially and temporally, contrarily to the FBL, and therefore its spatio-temporal variations can be seen as the major difference between the two methods.

The sequences were shown on a 46” LCD full HD resolution display with edge-lit (1.5D) LED backlight. The values of
the 16 LED segments and LC compensation values were first computed offline and then displayed on the platform through a special mode that allows controlling the backlight and LC values in real time. The participants were located at a distance of three times the display height and the display was tilted by an angle of 15° to represent a situation where several persons are watching the same display. This viewing angle also increases the leakage defect.

The three levels chosen for the peak white (i.e. the maximum displayed luminance) of the display were 75 Cd/m², 200 Cd/m² and 490 Cd/m² (which is the maximum of the platform). They were chosen to represent the range of possible realistic viewing peak white by the authors. The different peak whites were rendered by scaling the backlight to the desired value (linearity between the LED command and the obtained physical luminance was verified through measures).

The three ambient light levels were no light (approximately 0 lux), low ambient light (approximately 5 lux as measured on the display, 2400K) and higher ambient light (60 lux, 2800K). Those ambient light levels were obtained by varying the power supplied to three halogen lamps located symmetrically on each side of the display. This light was diffused with translucent paper. Participants were given time (a minimum of 5 minutes) to adapt to each light condition, i.e. the level of ambient light, before performing the required task.

10 stimuli were shown to the participants at each ambient light and peak white levels. The participants were instructed to evaluate the quality of each stimulus by placing a cursor on a continuous scale without label but with an indication of direction of increasing quality. A continuous scale was chosen in order to obtain data from an interval scale and not a category scale. Indeed with a division of the scale in 11 bins, theory shows that we are close to interval scale performance [25] and we avoid the quantization effect of the 5-points scale [26]. The algorithm order was randomized and the ambient light level, peak white level and sequence order followed Latin Square designs. The ambient light level was a block variable, meaning that the participants were rating all stimuli (5 sequences × 2 algorithms × 3 peak whites) at one lighting condition at one go, then they would get adapted to the next lighting condition and rate all the stimuli again and so on. Therefore the results from the experiment do not constitute a direct comparison of the quality at different ambient light levels. Indeed the participants need some time to accommodate to each ambient light condition, so it would have been impossible to link together quickly the quality assessments under different light conditions. Also the instructions to the participants did not mention this comparison either, so their grades are not calibrated across the different ambient light levels. Each session, i.e. each set of one ambient light level, lasted between 10-20 min.

20 observers participated in the experiment, even though they were naive regarding the purpose of the test, they are Bang & Olufsen employees and as such can be considered experienced in image processing. The experiment room has dark walls and is completely isolated from external lighting.

### IV. Results Analysis

In this section the statistical analysis of the subjective data is performed, first through an ANalysis Of Variance (ANOVA) of the results in Sec. IV-A and then by Partial Least Square (PLS) Regression between the subjective data and a set of subjective features in Sec. IV-B.

#### A. ANOVA analysis of the experiment

An ANOVA was performed on the subjective results with the Sequence, Algorithm, Peak White and Ambient Light as fixed factors and the Subjects as random factor, the statistical model included the main factors and the second and third order interactions. The ratings were measured in cm from the starting point of the scale, which was 10 cm long so ratings range from 0 to 10. The significant factors and interactions ($p < 0.05$), ranked by F-values, are shown in Table I for factors having an F value above 4.5 (a quarter of the biggest F value). The respective F value and degrees of freedom are also shown in this table.

<table>
<thead>
<tr>
<th>Factor</th>
<th>p-value</th>
<th>F</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm × Sequence</td>
<td>&lt;0.001</td>
<td>18.7</td>
<td>4</td>
</tr>
<tr>
<td>Algorithm × Sequence × Ambient light</td>
<td>&lt;0.001</td>
<td>9.1</td>
<td>8</td>
</tr>
<tr>
<td>Ambient light × Peak White</td>
<td>&lt;0.001</td>
<td>7.4</td>
<td>4</td>
</tr>
<tr>
<td>Sequence × Algorithm × Subject</td>
<td>&lt;0.001</td>
<td>6.7</td>
<td>76</td>
</tr>
<tr>
<td>Sequence</td>
<td>&lt;0.001</td>
<td>6.1</td>
<td>4</td>
</tr>
<tr>
<td>Algorithm × Peak White</td>
<td>0.006</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Ambient light × Peak White × Subject</td>
<td>&lt;0.001</td>
<td>5.4</td>
<td>76</td>
</tr>
<tr>
<td>Sequence × Peak White</td>
<td>&lt;0.001</td>
<td>4.8</td>
<td>8</td>
</tr>
<tr>
<td>Ambient light</td>
<td>0.012</td>
<td>4.7</td>
<td>2</td>
</tr>
</tbody>
</table>

**TABLE I:** Significant factors for the ANOVA of the quality grades (and corresponding $p$, F values and degrees of freedom df) with a F value above 4.5

From this table, it is visible that the factors Ambient Light and Peak White do influence the perceived quality. Directly and through interactions for the Ambient Light and only through interactions for the Peak White. Therefore the data are valid to model the impact of those factors on the quality. It can also be noted that the subjective preferences vary depending on the participant as is indicated by the significant interaction of the factor Subject with Sequence × Algorithm and Ambient light × Peak White. This subject preference is not desirable in the sense that we are aiming at a model for a global observer, and therefore we will average across subjects in the PLS analysis, however it was already noted for local backlight dimming by [15].

The most influential factor is the interaction between Sequence and Algorithm. As illustrated in Fig. 2, the quality of the two algorithms is rated significantly different for each of the sequences. For four sequences the GD is preferred and for the *Titles* sequence the full backlight (FBL) is preferred. The sequence *Titles* consists of credits scrolling up on a black background. As most of each frame is uniformly black, the GD algorithm lights up the LED segments as the titles scroll...
up. The resulting spatio-temporal variations of the leakage are quite noticeable and perceived as annoying by many observers. For the other sequences, a previous study [24] indicates that the GD algorithm is preferred because it exhibits higher contrast than a conventional LCD (FBL).

Fig. 2: MOS score as a function of the Algorithm and the Sequence (with 95% Confidence Interval)

The interaction between Ambient Light and Peak White is shown in Fig. 3. Even though the Peak White factor does not directly have an effect on the quality ratings, confirming the study by Guterman et al. [8], its interaction with the Ambient Light shows significant variations. As one could expect, the higher the Ambient light level, the higher the preferred Peak White level.

Fig. 3: MOS score as a function of the Ambient Light and the Peak White (with 95% Confidence Interval)

Based on the participants preferences, we can define test cases for which the ambient light and peak white levels are adapted to one another. Among the levels tried, the best balance is achieved by a peak white of 75 Cd/m² with no ambient light, a peak white of 200 Cd/m² at 5 lux and a peak white of either 200 or 490 Cd/m² at 60 lux (Fig. 3).

B. Modeling the quality using features

A set of 21 different objective features were computed from the modeled displayed sequences in order to try to analyze their influence on the quality.

Firstly for each stimuli (i.e. each sequence with the two backlight dimming methods) at each peak white level, the backlight and luminance of each displayed pixel were computed using the model described in [23].

In order to model the influence of the ambient light on the perceived quality, the ambient light level needed to be included in the modeled stimuli. This was done through two different methods. First, we considered that the light arriving into the eye of the observer from any part of the display is composed of the light emitted by the display plus the light reflected by the display. Therefore the first method consists of adding the reflected ambient light to the emitted light in the physical domain. The modeled luminance \( y_{light}^i \) for pixel \( i \) is then:

\[
y_{light}^i = y_i + k \pi E_{amb},
\]

where \( y_i \) is the modeled rendered luminance, \( k \) is the reflection coefficient of the display (3% in this case) and \( E_{amb} \) is the ambient illuminance in lux. Then the second chosen approach is the model developed by Devlin et al. in [14] where they use the following transform to account for luminance modification due to ambient light:

\[
y_{light}^i = \begin{cases} 
\frac{y_i L^2}{(L-L_R)^2+y_i L_R} & \text{if } 0 \leq y_i \leq L \\
\frac{y_i L_R}{(m-L_R)^2+y_i L_R} & \text{if } L \leq y_i \leq m \\
(\frac{y_i}{m-L_R})^2+m L_R (y_i+m+L_R-2L) & \text{if } L \leq y_i \leq m \\
(\frac{y_i}{m-L_R})^2+m L_R (y_i+m+L_R-2L) & \text{if } m \leq y_i \leq L
\end{cases}
\]

where \( L_R \) is the reflected light, \( m \) is the maximum value taken by the image and \( L \) a threshold value of 20% of \( m \).

The modeled rendered frames were therefore modified with those two methods a priori to the feature computation.

All features were computed using the luminance (physical intensity) or luma (pixel value) representation of the frames, respectively (meaning that color was not taken into account in this study).

The spatial error due to the rendering was represented through the Mean Square Error (MSE), the MSE on the leaking and clipping pixels only and the PSNR between the input image and the modeled displayed pixel values. The number of leaking and clipping pixels were also included to represent the rendering error. The characteristics of the rendered images were further analyzed through Michelson’s contrast globally over the image and the maximum over each LED segment, plus
the loss of spatial variations. The temporal variations were represented through the Sum of Squared Differences (SSD), the Temporal Information (TI), the maximum backlight variation over each segment and a flicker measure based on Michelson’s contrast, computed both on the modeled frames and on the modeled backlight. A summary of the computed features is shown in Table II.

As the objective features are computed separately for each frame (or each set of two successive frames), 11 temporal pooling methods were applied to turn each set of instantaneous measures into single values representing the whole sequence:

- 5 types of average (over all frames, the worst 10%, the best 10%, the first 2 seconds and the last 3 seconds),
- Minkowski summation with powers 2, 3, 4 and 5,
- the low pass FIR described by Hamberg and DeRidder and the asymmetrical pooling introduced by Ninassi et al. in [27] which consists of adding the average of the feature and a term representing its variation over time that favors the distortion decrease (compared to distortion increase).

Among the total of 231 possible predictors (21 features × 11 poolings), a set of 7 predictors were iteratively chosen to obtain the optimal prediction using the Unscrambler®X software. The selected features are: SSD, TI, the maximum backlight variation (computed in terms of physical luminance), the number of clipping pixels, the averaged over the whole modeled frame and maximum over each segment contrast and the loss of spatial variations. The temporal pooling achieving the best results was the one presented by Hamberg and DeRidder. The method accounting best for the ambient light is the one presented by Devlin et al. confirming the finding from [15].

The obtained model has a $R^2$ of 0.48 (a correlation of 0.69) with 5 optimal factors for the prediction. The resulting prediction is shown in Fig. 4. This prediction shows that the model is able to separate the different sequences but also that the ratings for each sequence are split into clusters and not completely predicted.

Even though the correlation value is relatively low compared to those achieved in many other studies where subjective mean opinion scores are compared against objective quality metrics, one should keep in mind that here it is not only the quality of the input signal that is evaluated but more globally the quality of the rendered signal. To the best of our knowledge, this paper is the first example of quality assessment of the ambient light and peak white level and therefore comparison with results from literature on quality assessment is biased as the scope of this model is wider than the one of the usual quality metrics.

V. CONCLUSION

This paper investigates the quality of dark high contrast videos rendered on a display with local backlight dimming capacity at various ambient light and peak white levels through a subjective experiment. An analysis of the results (ANOVA) shows significant impact of the interactions of the Sequence and Algorithm but also of the Ambient light and Peak White. As the ambient light increases the preferred Peak White increases also and those preferences allows us to define balanced cases for the Peak white x Ambient Light levels. A model of the subjective quality is then presented using PLS regression and achieves a correlation of 0.68.

REFERENCES
TABLE II: List and properties of features computed.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Support</th>
<th>Declination</th>
<th>Max/avg over segments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image</strong></td>
<td>Local standard deviation</td>
<td>Modeled image</td>
<td>All pixels</td>
<td>Both</td>
</tr>
<tr>
<td></td>
<td>Number of degraded pixels</td>
<td>Modeled image</td>
<td>Leaking pixels, Clipping pixels</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>Modeled image</td>
<td>All pixels, Clipping pixels, Leaking pixels</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Luma</td>
<td>Modeled image</td>
<td>All pixels</td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>Loss of spatial variations</td>
<td>Modeled image</td>
<td>All pixels</td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>Contrast (Michelson [29])</td>
<td>Modeled image</td>
<td>All pixels</td>
<td>Both</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Backlight</strong></td>
<td>Maximum intensity</td>
<td>LED intensity</td>
<td>All pixels</td>
<td>Maximum</td>
</tr>
<tr>
<td></td>
<td>Power consumption</td>
<td>LED intensity</td>
<td></td>
<td>Average</td>
</tr>
<tr>
<td><strong>Temporal change</strong></td>
<td>Sum Squared Differences (SSD [27])</td>
<td>Modeled image</td>
<td>All pixels</td>
<td>Maximum</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Both</td>
</tr>
<tr>
<td></td>
<td>Temporal Information (TI [28])</td>
<td>Modeled image</td>
<td>All pixels, Leaking pixels, Clipping pixels</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Maximum Backlight variation</td>
<td>Modeled image</td>
<td>Backlight</td>
<td>Maximum</td>
</tr>
<tr>
<td></td>
<td>Luma variation</td>
<td>Modeled image</td>
<td>All pixels</td>
<td>Both</td>
</tr>
<tr>
<td></td>
<td>Flicker measure</td>
<td>Modeled image</td>
<td>All pixels</td>
<td>Both</td>
</tr>
<tr>
<td></td>
<td>Flicker measure</td>
<td>Modeled image</td>
<td>Backlight</td>
<td>N/A</td>
</tr>
</tbody>
</table>