Introduction

The driving task mainly relies upon visual information. Flight- and driving- simulators are designed to reproduce the perception of reality rather than the physics of the scene. So when it comes to using driving simulation for studying road visibility, care should be taken on visual cues. Unfortunately, even when a physically correct representation of the road environment is available, seldom can it be faithfully displayed. The luminance range that can be achieved by a simulator is restricted compared to the reality, which makes image compression unavoidable.

Real-world luminance levels need to be compressed in order to fit within the limited dynamic range of video monitors and projectors. This is achieved by a so-called “tone-mapping” operation. But the question arises whether tampering with the visual cue will not alter visual performance, possibly invalidating comparisons between behaviours observed in real and simulated conditions.

However, research is very active in the field of image display. Advanced tone-mapping techniques are usually designed on physiological grounds, to ensure visually satisfactory results. In this paper, we present a psychometric experiment for assessing the quality of tone-mapping in terms of visual appearance and visual performance. Subjects were asked to perform a specific task, both in a reference high dynamic range scene, and in tone-mapped images of this scene displayed on a CRT monitor. Four tone-mapping algorithms were tested, none of which fully succeeded in achieving the original look and feel.

Psychometric assessment

During the last two decades, several algorithms have been proposed for tone-mapping, each new proposal achieving a further degree of improvement. Starting with simple perceptual laws such as Weber’s law or Stevens’s law [1], authors have introduced control of the contrast threshold [2] or luminance histogram adjustment [3].

The question arises whether these algorithms achieve the goal they have been designed for, or better, the goal the user of the simulators wants to achieve.

The quality of the image transformation is not a purely subjective concept. It can be measured by estimating the correlation between either the visual appearance or the visual performance, before and after the tone-mapping transformation. The relevant criterion depends on the visual task which is to be performed, in a simulator, with the images.

In this paper, we propose to rate the merit of four models in achieving satisfactory tone-mapping, both on an appearance[4] and on a visual performance point of view. A reference scene is used, consisting of a wide wall receiving controlled illumination. A test picture is presented in the middle of the wall, embedded in an achromatic noise background. The test picture is different in the visual appearance and in the visual performance experiments.

The test picture for visual appearance uses the highest dynamic range. The test picture for visual performance includes just noticeable targets. The observer’s judgement refers to appearance or to discrimination.

Simulations of this scene were produced, using four different tone-mapping algorithms, and displayed on a CRT monitor. The visual appearance and performance are evaluated in the reference scene and in four simulated situations by the same observers.

In both cases, the hypothesis is that a tone-mapping algorithm which performs well should yield the same results in the reference scene and in the simulated scene.

Tone-mapping algorithms

Four tone-mapping algorithms were implemented [5], three of which are linear. It was decided to focus on linear procedures, because it is the only way to keep the same performance distortion on the whole scene (assuming that the contrast detection can be linked with a $\Delta L/L$ factor [6]). Every algorithm transposes the original scene dynamics in its own way.

Algorithm 1

Algorithm 1 (maximum) consists in mapping the whole luminance range into the display range:

$$L^{\mathrm{d}} = \frac{L^{\mathrm{d}}_{\max}}{L_{\max}} L \quad (1)$$

where L^{d} is the displayed luminance, L is the luminance of the reference scene at the corresponding location, $L^{\mathrm{d}}{}_{\max}$ is the maximum display luminance available, and $L_{\max}$ is the maximum luminance in the scene.

Algorithm 2

Algorithm 2 (mean) compensates for the high sensitivity of Algorithm 1 regarding a single spot value. Instead of mapping the maximum luminance value to the maximum display value, the mean value $\langle L \rangle$ is mapped to half the maximum display value:

$$L^{\mathrm{d}} = \frac{L^{\mathrm{d}}_{\max}}{2\langle L \rangle} L \quad (2)$$

Algorithm 3

Algorithm 3 (Ward [2]) introduces the visual sensitivity of the human eye, in order to respect the visual performance. The adaptation luminance in the real scene L_{a} is computed, as well as the corresponding sensitivity threshold $\Delta L_{\mathrm{t}}(L_{\mathrm{a}})$. The adaptation luminance generated by the display is assumed to be half the maximum display luminance. Then, the slope of the linear mapping is computed in order to get the same visibility level:

$$L^{\mathrm{d}} = \frac{\Delta L_{\mathrm{t}}\left(L^{\mathrm{d}}_{\max}/2\right)}{\Delta L_{\mathrm{t}}\left(L_{\mathrm{a}}\right)} L \quad (3)$$

The expression of ΔL_{t} is computed from experimental data [6].

Algorithm 4

Algorithm 4 (histogram [3]) is not linear. Nevertheless, it is often used in tone-mapping applications, and leads to qualitatively good results. Its purpose is to optimize the luminance histogram, in order to make maximum use of the display range.

The visual appearance experiment

Reference scene

We have built the real scene in a room of the Vision laboratory of the *Museum National d'Histoire Naturelle*. It consists of a wall including surfaces that receive controlled illumination and reflect light in the whole room in a diffuse mode.

The test consists of two different horizontal gradations of gray levels embedded in an achromatic noise background of high spatial frequency. The test is printed on paper and directly illuminated at a 45 deg angle by a metal halide overhead projector (6250 K). The periphery consists of a unique achromatic noise of medium spatial frequency obtained by projecting a transparency on the white wall. The observer faces the wall at a 2 m distance and views all surfaces in a natural way.

Although the dynamics of the gradation is always maximum, the gamma of the gradation could be set at one out of six fixed values, and was changed from one trial to another. Plates have been manufactured with

every possible pair of different gradations. Three series of plates were mounted on a drum and presented in a random sequence to the observer who was invited to choose the one he found the more “regular”. Seven observers have participated in the experiment, assessing 30 comparison judgments for each pair of gamma, in 10 sessions.

Calibration was carried out *in situ*. In particular, stray light was accounted for, as it greatly modifies contrast ratios. To calibrate the test and to control the printing process on each plate that was included in the real scene, we produced two posterised gradations on the same print as the graded arrangement. The gamma of each gradation was computed using the position of every border of the posterised gradation as well as the luminance factor of each uniform area of the posterised gradation. The luminance factor was measured using a Minolta L100 photometer. Gamma values computed from the measurements were sorted out and only the plates that fit within six restricted classes of gamma were used in the experiment. The transparency that produces the noise at the periphery has been printed with square elements, the density of which was controlled by digital code values selected from a series of 10 values. The transparency has been emptied at the central position containing the test with the gradations, and at the positions of four square elements in the periphery. The luminance of the patches of the noisy periphery was measured *in situ*. Luminance values were distributed between 44.7 and 793 $cd.m^{-2}$.

Simulated scene

The simulation was produced on a CRT display (Fig. 1). Images were created in which each element – gradations, background noise, periphery noise – had the same angular dimension as in the real scene, when they were viewed on the CRT display at a 60 cm distance. The CRT display was calibrated following the Gain-Offset-Gamma method recommended by the CIE [7].

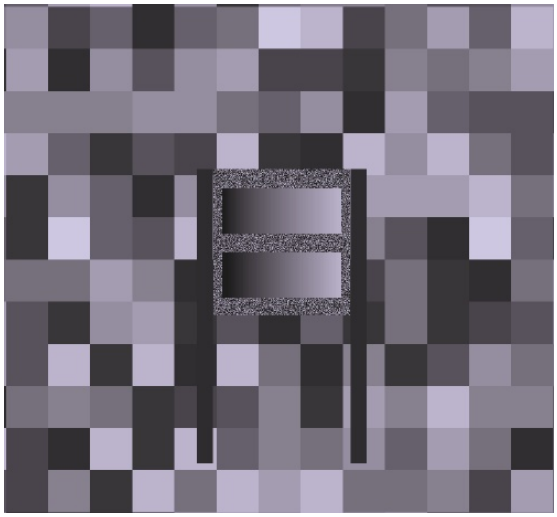

Figure 1. Gradations, detail of the simulated scene.

Pseudo-random sequences of images were prepared for each algorithm. Each pair of simulated gradations, differing by 1 or 2 gamma steps, was presented twice in a sequence. Each observer performed 15 sessions consisting of one sequence for every algorithms.

Results

Optimal gamma for the reference gradations

The results of the experiment with the reference scene show that observers are able to judge accurately which gradation is the best representative of the optimal gamma. A count was made of the number of occurrences of each gamma as preferred by the observer. Under examination of the distribution of preferred choice, the rating of gamma values is about symmetric around the optimal gamma, on a logarithmic gamma scale. Indeed, for 6 observers out of 7, the distribution of occurrences shows a clear maximum which can be

modeled by a third order equation with the logarithmic value of the gamma as variable (Fig. 2). This leads to the determination of the preferred gamma value in the reality (Tab. 1).

For the seventh observer, the distribution is monotonic within the range of gamma values presented but the slope indicates that an optimal gamma would probably have been found if lower gamma values had been presented.

Table 1. Preferred gamma values for gradations presented in the real scene.

Obs1	Obs2	obs3	Obs4	Obs5	Obs6	Obs7
1.837	1.827	1.511	1.677	1.505	1.396	< 1.14

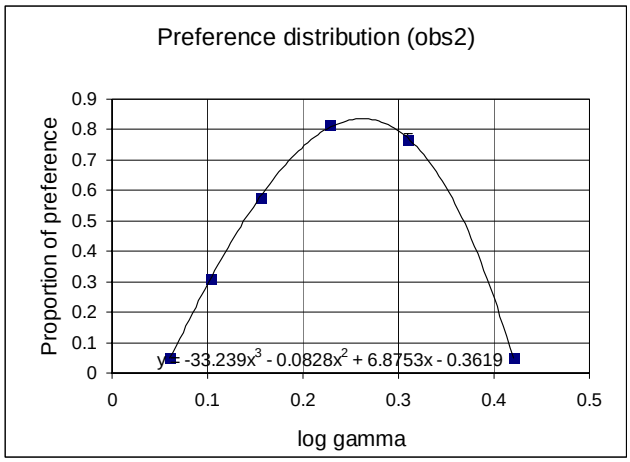

Figure 2. Distribution of occurrences of each gamma as preferred by one observer.

The question arises whether the inter-observer variability reflects differences in scaling ability or differences in interpreting the instructions. Indeed, some observers have reported that they would judge differently the smoothness or the balance of the gradation. Eventually, every observers had to decide upon their criterion.

Rating the algorithm for appearance

The hypothesis is that a tone-mapping algorithm which performs well should yield the same optimal gamma as in the reality. However, if the simulated optimal gamma value is lower than the real optimal gamma, it means that the algorithm produces gamma values higher than predicted. Conversely for the opposite result.

Our results show that none of the algorithms that have been tested perform well, especially when the optimal gamma is not included within the range of gamma values that have been simulated. For 3 algorithms out of 4 (“maximum”, “mean” and “Ward”), the optimal gamma would fall beyond the higher boundary for 2 observers out of 7. For the other algorithm (“histogram”), the optimal gamma would fall below the lower boundary for 3 observers out of 7 (Fig. 3).

Rather than averaging the choice of the observers, we have compared, for each observer and each simulation, the ratio between the optimal gamma given in the simulated situation and the optimal gamma given in the real situation, in order to discount the inter-observer variability (Fig. 3).

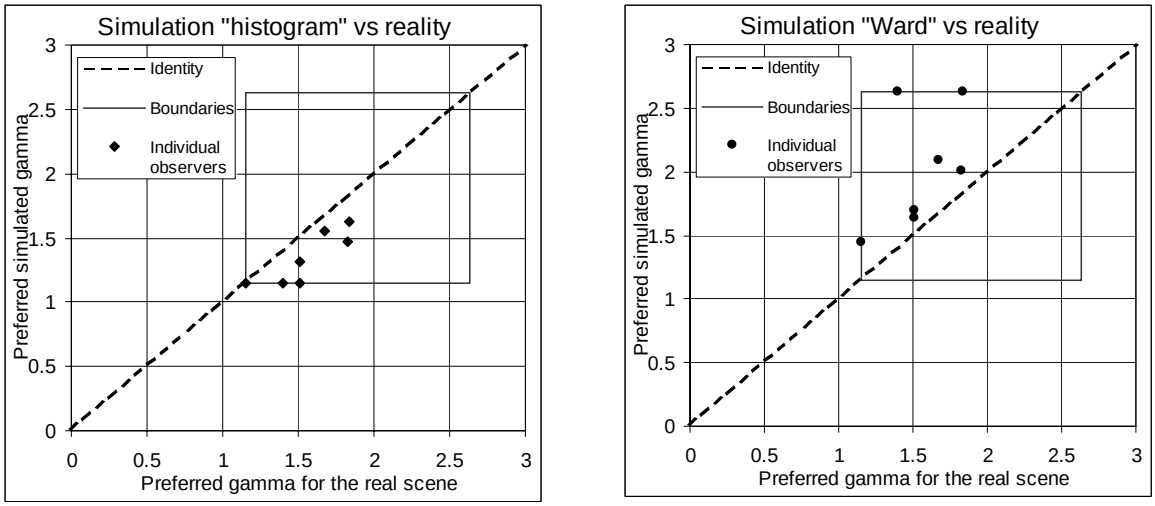

Figure 3. Preferred simulated gamma vs preferred gamma for the reference scene for two algorithms.

It is worth noting that the 7th observer who could not find his optimal gamma within the range proposed for real scenes has been able to find it for algorithms 2, 3 and 4, where other observers had failed. This confirms the necessity to take into account individual preferences and supports our decision to compare results individually.

The visual performance experiment

Reference scene

The same tools and methodology as in the appearance experiment were used for the visual performance rating, except for the test area.

For assessing the visual performance, a discrimination task was proposed. The test consists of a set of 4 lines printed on a uniform grey background. One of the line was different (three continuous lines and one dashed, or three dashed lines and one continuous) (Fig. 4). Observers were asked to point out the odd one. The answer delay was recorded, and then taken into account in the visual performance evaluation, according to CIE TC 1-19 proposal (accuracy × speed) [8]. The contrast of the 4 lines was selected among 6 values, 3 positive and 3 negative contrast values (Table 2). Every possible combination of contrast and line configuration has been constructed (3 positive contrast values and 3 negative contrast values, odd continuous or odd dashed line, 4 possible positions), ending into 48 plates which were mounted on the drum and presented in a random sequence to the observer. Hence, the same contrast was presented in 8 plates, within each run. After 5 runs, the observer discrimination ability has been screened 40 times on each contrast. 5 observers have served on the performance experiment, in the reality as in the simulation.

The contrast of the lines on the discrimination plates was computed from *in situ* photometric measurements recorded within a line and at the border of the line, using a highly sensitive photometer (Spectra Pritchard from Photo Research). We used the Michelson contrast $C = (L_f - L_b)/(L_f + L_b)$, L_b being the background luminance and L_f the foreground luminance.

-0.0219	-0.0140	-0.0069	0.00680	0.0163	0.0240

*Table 2. Average contrast values of the lines presented in the real scene (*in situ *measurements included stray light).*

Simulated scene

We used the same four tone-mapping algorithms as for the visual appearance experiment to rate the visual performance of the observer with the simulated scene. Images differ from the previous series only in the test area (Fig. 4). All contrast values were transposed using the previously described algorithms. A series of 24 images,

showing 4 times the same transposed contrast arranged with different line configurations was prepared. Each observer performed 10 sessions and tested all algorithms.

Figure 4. Lines, detail of the simulated scene.

As the tone-mapping procedure modifies the luminance distribution in the scene, the contrast values that were presented in the simulation were different from the contrast values of the original scene. Measurement of the contrast of the simulated scene were performed to check the reliability of the CRT calibration. Note that the contrast of the image that is of interest to our experiment is the contrast which is intended to be transposed, not the contrast which is actually displayed.

Results

Visual performance in the reference situation

For the lowest contrasts that were presented in the reality, the rating is close to chance because the lines were hardly visible. The medium contrast was almost always discriminated correctly (Fig. 5). Nevertheless the discrimination of the medium contrast was difficult and required longer time than for the highest contrast (Fig. 6). So instead of analysing the rating of good responses, we have computed the "Visual performance" as defined by the CIE:

$$\textit{Visual performance (VP)} = \textit{accuracy} \times \textit{speed}$$

or

$$VP = \textit{proportion of good responses / delay of the response}$$

Therefore, visual performance is good (close to 1) for the highest contrast because the proportion of good response is close to 100 % and because the observer responds rapidly (within 1 second). It is very poor (close to zero) for the lowest contrast because the delay of the response is very long. Visual performance declines smoothly in between. Using visual performance as an indicator proves to be efficient to smoothen the results (Fig. 7).

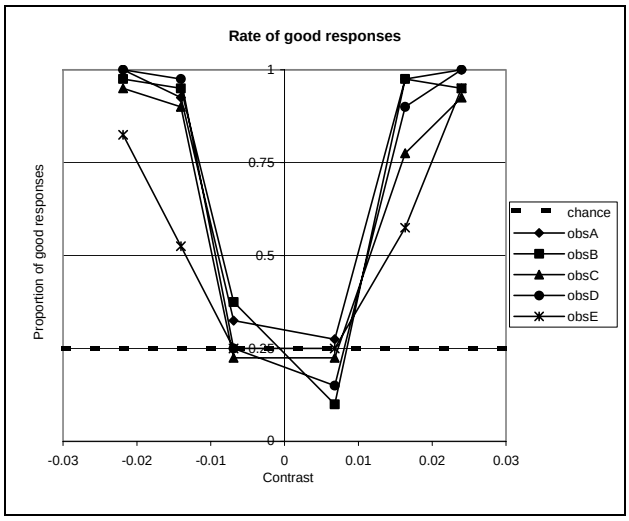

Figure 5. Rate of good responses vs contrast for the discrimination task in the reference situation. 0.25 is the chance level.

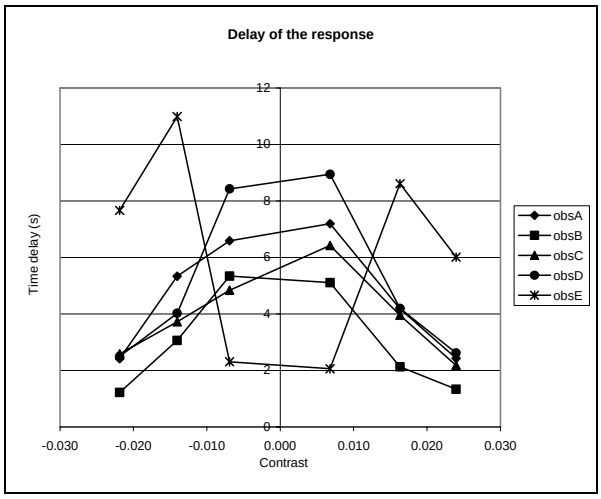

Figure 6. Delay of the response vs contrast for the discrimination task in the reference situation. Observer E has responded very rapidly to the lowest contrast which he could not see at all.

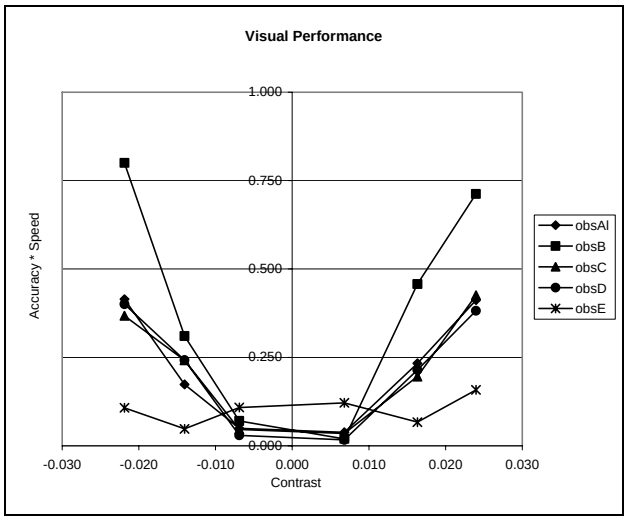

Figure 7. Visual performance vs contrast for the discrimination task.

Rating the algorithm for discrimination

Visual performance is used as the index to compare the simulation and the reality. The hypothesis is that an algorithm that performs well should yield the same visual performance index as in the reality. Our results show that it is not the case.

For 3 algorithms out of 4, the visual performance falls down to a very low value, even for the highest contrast which the algorithm should have transposed. Indeed, the photometric measurements indicate that all contrast values are within the range [-0.016, +0.016], but the lines have been displayed at the same angular size as in the reality, so they only spread over 3 pixels and are hardly visible. Yet, observers have complained about the difficulty of the task. For algorithm "histo", all contrasts were clearly distinguishable, except for the smallest which was transposed to a null value. Indeed the photometric measurement has revealed that negative contrast values were transposed to – 0.06 and positive contrast values to + 0.11. The task was, thus, very easy for the observers who succeeded within 1 second (Fig. 8).

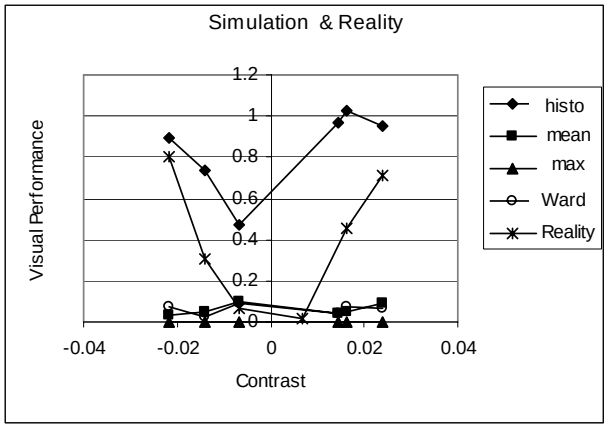

Figure 8. Comparison between the simulation and the reality for one observer. Other observers yield similar responses. (Due to an error in the procedure, one contrast value in the simulation corresponded to a real contrast value different from what it should have been. This is visible on the abscissa scale of the figure which has been plotted correctly.)

None of the tested algorithms manage in reproducing the visual performance of the reference scene, not even the Ward algorithm, which was designed for this purpose.

Discussion

We observe two classes of algorithms. On the one hand the "histo" algorithm, which is non-linear, distorts the low level contrast values. On the other hand the three linear algorithms do not allow the observers to perform correctly the discrimination. The contrast domain in the reference and displayed scenes is between –0.025 and +0.025, but contrast sensitivity of the human eye is much smaller around the displayed luminance level. This suggests that testing either a wider contrast domain or contrast on large patches, might yield curves that have the same shape with linear algorithms as in the reference scene.

Conclusion

After our experiments, it appears that none of the tone-mapping algorithms that have been tested represent the reality. The four algorithms fall in two classes, either under- or over-estimating the gamma values. Despite inter-observer variability, observers agree on their judgment. The same distinction in two classes, under-estimating and over-estimating the visual performance, is found with the visual performance task. The same algorithms are gathered in the same two classes in both experiments.

This experiment is quite hard, but seems necessary if one wants to extend the results obtained on a driving simulator (visual performance, visual appearance) to a real driving situation, including the real luminous environment.

This work suggests some ideas to build a specific tone-mapping algorithm, which should be defined in order to match specifically these kinds of psychometric experiments, and reproduce as much as possible the visual performances of a driver.

Acknowledgements

This research is part of the VOIR project, supported by the French Ministry of Research and associating OKTAL, the LCPC, the INRETS and the CNRS/CEPA. We thank the observers for assisting in the experiment.

References

1. J. Tumblin and H. Rushmeier, “Tone reproduction for Realistic Images”, *IEEE Computer Graphics & Applications*, **13**(6), November 1993, pp. 42–48.
2. G. Ward, “A Contrast-based Scale Factor for Luminance Display”, in *Graphics Gems IV*, ed. P. S. Heckbert, 1994, pp. 391–397.
3. G. W. Larson, H. Rushmeier and C. Piatko, “A Visibility Matching Tone Reproduction Operator for High Dymanic Range Scenes”, *IEEE Transactions on Visualization and Computer Graphics*, **3**(4), October-December 1997, pp. 291–306.
4. F. Viénot, C. Boust, R. Brémond, E. Dumont, rating gradations for tone-mapping algorithms, *IS&T, CGIV 2002, Poitiers, 2-5 April 2002,* pp. 221-225.
5. G. Pouliquen, “Respect des niveaux de visibilité dans la restitution d’images de synthèse”. Rapport de DEA, ESME/LCPC, septembre 1999.
6. Y. Le Grand, “Optique physiologique – Tome 2 : Lumière et couleurs”, Masson et Cie, Paris, France, 1972 (2nd edition).
7. CIE, “The relationship between digital and colorimetric data for controlled CRT displays”, CIE publication 122-1996.
8. CIE, “The correlation of models for vision and visual performance”, CIE Publ 145-2002.