Towards a perceptual quality metric for computer-generated images

Pierre Boulenguez, Boris Airieau, Mohamed-Chaker Larabi, Daniel Meneveaux

Towards a perceptual quality metric for computer-generated images

Pierre Boulenguez, Boris Airieau, Mohamed-Chaker Larabi, Daniel Meneveaux

XLIM-SIC Labs, Poitiers University, France

ABSTRACT

The physical validation of computer-generated images (CGIs) has received a lot of attention from the computer graphics community, as opposed to the assessment of these images' psychovisual quality. The field indeed lacks the proper tools to quantify the perceptual quality of a CGI; this paper engages in the construction of such a metric. A psychovisual experiment was submitted to a representative panel of observers, where participants were asked to score the overall quality and aspects of this quality for several CGIs. An analytical quality model, fit to the data, next gives insight into the relative perceptual importances of these aspects. Accuracy in the simulation of shadows, good contrast and absence of noise were found to have a major impact on the perceived quality, rather than precise anti-aliasing and faithfull color bleeding.

Keywords: Quality metric, computer-generated images, global illumination, psychovisual evaluations, statistical analysis.

1. INTRODUCTION

Realistic computer-generated images (CGIs) have emerged as a key component to such wide-ranging fields as architecture, road traffic, tunnel lighting, archeology, cultural heritage and electronic gaming. These images originate from the vibrant discipline of computer graphics realistic rendering, of which the ultimate goal can be seen as the simulation of every light related phenomena occuring within a given virtual environment.

In view of the theoretical and computational complexity of the task at hand, numerous approaches have specialized in the simulation of *diffuse* light scattering.[1,2] Phenomena induced by non-diffuse light/matter interactions (*e.g.* gloss, specularity, transparency, caustics, anisotropy) are then rendered separately.[3]

A number of studies also have addressed the problem of reducing complexity based on perceptual considerations,[4–6] but the field still lacks a dedicated perceptual quality metric adapted to CGIs. Such a tool would be beneficial to interactive-time rendering, allowing for a focusing on the most perceptually significant phenomena, as well as to off-line simulations, where the stop criterion of the iterative light distribution could be based on perceptual considerations, rather than purely empirical. Comparable perceptually-driven approaches have been employed with great success for speech, audio and video compression.[7–9]

This paper engages in the construction of such a perceptual quality metric adapted to CGIs. First, the work context at the frontier between rendering and perception is introduced (Section 2). Five criteria of the psychovisual quality of CGIs are then detailed (Section 3). These criteria were assessed by observers through a psychovisual experiment, described along with the statistical tools involved (Section 4). Results obtained are discussed (Section 5). This paper ends with conclusions and perspectives for future research.

E-mail: pierre.boulenguez@univ-poitiers.fr, boris.airieau@univ-poitiers.fr, larabi@sic.univ-poitiers.fr, daniel@sic.univ-poitiers.fr

2. RELATED WORK

2.1 Work context in rendering

The problem of computing the light distribution within a virtual environment is formalized by *the rendering equation*.[10] Recursive by nature, it cannot be solved analytically for arbitrary environments and its numerical approximation by realistic rendering approaches is referred to as *global illumination*.

Monte Carlo integration can be seen as the reference methods for offline simulations. Path-tracing,[10] for instance, estimates radiance impinging on each pixel, following random paths from the camera towards the sources. The approach, although unbiased, results in noisy CGIs unless a considerable number of paths are computed. Refinements such as importance sampling,[11] bidirectional path tracing,[12] or metropolis methods[13,14] enhance convergence. Irradiance cache[15,16] also reduces computational complexity, using interpolation between accurate samples. They are distributed in the environment and can be reused to compute new images from nearby camera positions. Interpolation results in low-frequency noise and an adequate choice of interpolation parameters is not straightforward.[17]

Photon-tracing approaches,[3] on the other hand, construct paths directly from the sources, and simulations are independent from the camera position. The impacts of the paths within the environment are represented as photons, stored in a so-called photon-map. Photon-map-based operators are inherently biased and the resulting CGIs often exhibit low-frequency noise. A final gather stage circumvents the problem but, in turn, introduces high-frequency noise. Photon streaming[18] avoids noise flickering for interactive-time rendering in dynamic environments. Finally, instant-radiosity methods[19] represent impacts as virtual point light sources. These methods are unbiased but result in check-pattern-like noise due to multiple shadow casting.

For interactive-time rendering, a compromise has to be found between physical plausibility and severe time constraints. Historically, the ambient term has been seen as the most practical approximation for diffuse lighting. Ambient occlusion[20] can modulate this ambient term using an estimation of occluded light at each viewed position. Finally, environment gather[21] stores light scattered in an environment map.

2.2 Work context in perception

The link between perception and rendering is not new. The early years of CGIs were indeed essentially perceptually driven, as the validation of a particular approach was carried out by visual inspection. The orientation towards physically based rendering has lessened the importance of visual assessment for offline simulations, but it remains strong for interactive-time rendering.

In spite of this fundamental relationship, the first formal psychovisual evaluation of CGIs was, to our knowledge, only carried out in a study[22] where participants were asked to assess the implications of various visual cues on the rendering of depth. Other experiments were subsequently conducted[23] to evaluate the influence of shadow rendering strategies on the perception of object shape. This research led to the important and somewhat confusing conclusion that less physically accurate shadows may be preferable in some tasks requiring unambiguous perception of the shape.

Another pioneering study[4] concerns the perceptual implications of environment simplifications (*e.g.* geometry, textures, shading models). An experiment[24] was, for instance, recently conducted to assess the psychovisual impact of approximate indirect visibility computation and notably their impact on shadow realism.

A number of studies[5,6] have also engaged in *selected rendering*, which refers to the ability of a given algorithm to deactivate certain features of the rendering process (*e.g.* direct lighting, indirect diffuse lighting, mirror reflections, glossy reflections, transparency). The perceptual importance of each stage of the rendering process can then assessed by formal psychovisual experiments. In such an experiment,[5] for instance, ten participants were asked to assess the quality of CGIs with indirect diffuse, glossy and direct lighting alternatively activated.

The aformentioned studies aim at reducing the rendering time of a given algorithm while maintaining an acceptable level of perceived quality. To our knowledge, no formal inter-algorithms psychovisual evaluations were conducted. For instance, there is no way to know how, for the diffuse lighting component, the empirical ambient term compares to a physically accurate global illumination simulation, in terms of perceptual quality.

3. A SET OF CGI'S PSYCHOVISUAL QUALITY CRITERIA

Let $\{C_1, \ldots, C_n\}$ be a set of criteria and Q the overall pyschovisual quality of a CGI. The ultimate goal of this research is to find an optimal function f such that:

$$Q = f(C_1, \ldots, C_n). \tag{1}$$

Ideally, the set $\{C_1, \ldots, C_n\}$ would encompass every possible aspect of a CGI's psychovisual quality. Such a set would obviously have exceedingly large cardinality, and contain criteria that might be *fuzzy* or non-intuitive. Yet, as each criterion has to be explained to and assessed by participants during the psychovisual experiment, a first set of five criteria, empirically known as important in diffuse environments, was constructed. Those criteria are: color bleeding C_b, shadow C_s, noise C_n, aliasing C_a and contrast C_c. They are described in the remainder of this section.

3.1 Color Bleeding C_b

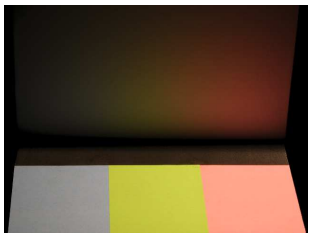


Figure 1: Illustration of the color bleeding criterion (real case).

Color bleeding, denoted C_b, is a subtle phenomenon by which the light reflected by a colored surface becomes *tinted* by the color of that surface. Figure 1 illustrates the effect with colored paper sheets and a diffuse white plate. Closely related to the interreflections in the environment, the phenomenon is difficult to render faithfully without a complete global illumination simulation.

3.2 Shadows C_s

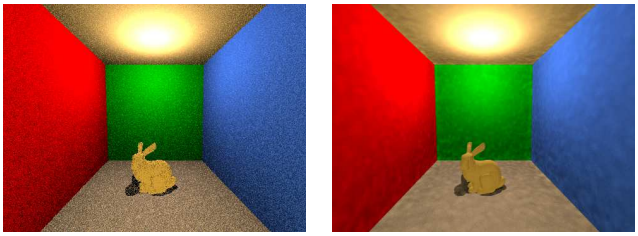
(a) Flat shadow. (b) Gradient shadow.

Figure 2: Illustration of the shadows criterion.

Shadows are another key challenge as their appearance has already been identified as critical to a CGI's perceptual quality. The luminosity gradient across a shadow, which tends to be darker in the vicinity of the occulting object, is a particularly difficult aspect to render. Figure 2 illustrates the phenomenon.

3.3 Noise C_n

The noise criterion, illustrated in Figure 3, can be distinguished into high and low frequency noise. As stated in Section 2.1, many stochastic rendering approaches, despite accurate global light distribution, are also prone to produce noise.



(a) High frequency noise. (b) Low frequency noise.

Figure 3: Illustration of the noise criterion.

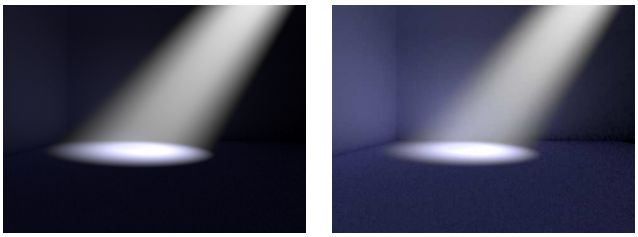
(a) Aliased rendering. (b) Anti-aliased rendering.

Figure 4: Illustration of the aliasing criterion.

3.4 Aliasing C_a

Edge aliasing is a common issue with some rendering methods. Figure 4a details a jagged junction between two walls in the background, while Figure 4b shows the same image with anti-aliasing.

3.5 Contrast C_c



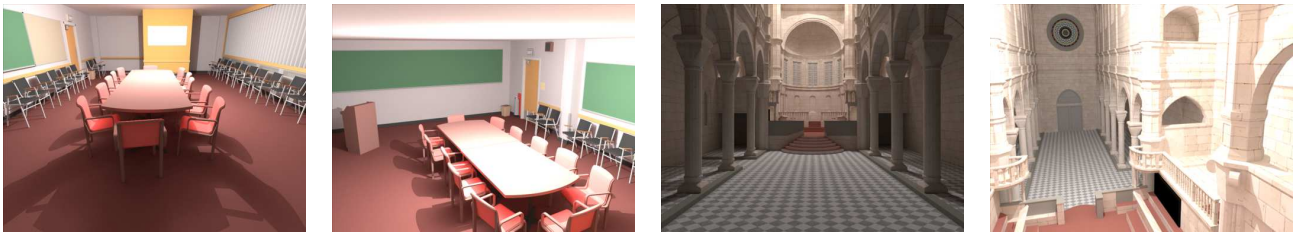
(a) Uneasy contrast. (b) Comfortable contrast.

Figure 5: Incidence of the contrast criterion.

Contrast and luminosity are well known attributes of an image. In the context of CGIs, they are dependent on both the main light distribution process, as well as on the quantification stage (or tone mapping), that converts high dynamic range images into standard 8-bit per component images. Figure 5 presents renderings of the same environment with varying contrast and luminosity settings.

4. SUBJECTIVE QUALITY EXPERIMENT

To evaluate the influence of the criteria $\{C_b, C_s, C_n, C_a, C_c\}$ on the overall perceived quality Q, a subjective quality experiment was constructed and carried out. This section describes the procedure in detail.



(a) Conference I. (b) Conference II. (c) Sibenik I. (d) Sibenik II.

(e) Sponza I. (f) Sponza II. (g) Cornell Box.

Figure 6: The seven views and corresponding scenes used for the experiment. The whole data set (49 CGIs) was made available online.[25]

4.1 Construction

4.1.1 Test set

Seven rendering algorithms that produce the effects described in Section 3 were selected: ambient term (AT), ambient occlusion[20] (AO), environment gather[21] (EG), photon streaming[18] (PS), photon mapping[3] (PM), photon mapping with a final gather pass[3] (FG) and Metropolis light transport[14] (ML). The latter is the most computationally intensive and is considered the most physically plausible.

Seven views from four widely used scenes in computer graphics (see Figure 6) were also chosen. They represent a suitable variability of content for the intended experiment and are freely available. Each scene has been used as input to the aforementioned algorithms in order to generate the subjective experiment stimuli.

4.1.2 Algorithms parametrization

The ML algorithm has been tuned to provide the most accurate rendering, independent of any computational complexity considerations. The other methods were voluntarily used with degraded parameters in order to highlight each criteria.

Precisely, PM and FG CGIs were generated by tracing 200,000 initial photons and storing 1,000,000 photons in the photon-map. The final gather pass was done with 32 rays. AT values have been experimentally set for each scene. AO images were generated with ray-tracing using 64 rays. PS images were generated with 1,024 virtual light sources and 120,000 photons stored. For EG, 128 rays of final gather have been used.

Regarding computation time, PM images were rendered in a few seconds, while their FG counterparts needed several minutes. All other methods only needed a few milliseconds to compute.

4.1.3 Test environment

The experiment was conducted in a dedicated room constructed following the ITU recommendations[9] and equipped with a 30" LCD monitor with a native resolution of 2,560 $\times$ 1,600 pixels. The ratio of inactive screen luminance to peak luminance was kept below a value of 0.02. Lighting was ensured using 4 controlled neon tubes delivering a D50 light, and oriented to obtain 64 lux at the display while avoiding direct illumination. The calibration of the screen was performed using a Gretag Macbeth EyeOne calibration device. The viewing distance was set at 4 times the height of the presented images.

4.2 Overall test procedure

4.2.1 Participants

Thirty non-CG-experts participated in the experiment. This number complies with the minimum requirement,[9] even if some outliers are rejected.

The panel of participants has been characterized in terms of gender, occupation and age, but also in terms of knowledge in computer science and computer graphics. In particular, 21 males and 9 females participated. They were, overall, university staff and students, and their ages were distributed in the range of 23 to 51 years with the average age being 31.4 years.

4.2.2 Stimuli presentation

The experiment was constructed as a three-staged procedure: Initially, background information on the participant was gathered. Notably, visual acuity and color blindness were checked (using FrACT[26] and Ishihara plates) and a small quiz waz administered (see questionnaires[25]).

Next, the evaluation procedure and the significance of each criterion were explained to the participant. A normalized speech illustrated by slides was used. The participant was invited to ask questions if needed and had access to the slides during the evaluation stage.

Finally, the evaluation *per se* consisted in presenting the 49 CGIs in a fully random order (in order to average training bias). For each CGI, the participant was asked to assess its overall quality Q and the five criteria $\{C_b, C_s, C_n, C_a, C_c\}$ (*cf.* Section 3). As there is no reference in the field (although ML can be seen as the most physically accurate), the selected protocol is a single-stimulus one without time limitation. The scoring was done on a continuous scale without numbers but divided in an odd number of sections with two quality adjectives (poor and excellent).

4.3 Statistical analysis

4.3.1 MOS and confidence intervals

The Mean Opinion Score (MOS) $\bar{u}_k$ was computed for each presentation:

$$\bar{u}_k = \frac{1}{N} \sum_{i=1}^{N} u_{ik}, \tag{2}$$

where u_{ik} denotes the score of the i^{th} observer for the k^{th} image and $N = 30$ is the number of participants. The confidence interval associated with the MOS is given by:

$$\lfloor \bar{u}_k - \delta_k, \bar{u}_k + \delta_k \rfloor. \tag{3}$$

The deviation term δ_k is given for a 95% confidence interval, calculated as follows:[9]

$$\delta_k = 1.96 \ \frac{\sigma_k}{\sqrt{N}}, \tag{4}$$

where σ_k is the standard deviation.

One of the objectives of the statistical analysis was to be able to eliminate from the final results either a particular score or observer (outlier). In this paper, the method based on the Kurtosis coefficient[9] was employed.

4.3.2 Analytical modeling

Although the construction of an optimal quality function f (*cf.* Equation 1) is the ultimate goal of this research, as a first step, a linear regression was applied to establish a hierarchy in the contribution of each criterion on the overall perceived quality Q:

$$Q = \beta_b C_b + \beta_s C_s + \beta_n C_n + \beta_a C_a + \beta_c C_c + \beta_0, \tag{5}$$

where $\{\beta_b, \beta_s, \beta_n, \beta_a, \beta_c\}$ are the independent contributions of the criteria $\{C_b, C_s, C_n, C_a, C_c\}$ and β_0 is the intercept term.

5. RESULTS AND DISCUSSION

Figure 7 provides Mean Opinion Scores (MOS) and associated intervals for the seven view/algorithm combinations (as well as the average over the whole data set). The scores are presented with radar graphs in order to ease their interpretation. Figure 7h shows that noise C_n was easy to score because the values are spread over almost the entire scale range. As expected, ML obtained the highest scores for all the psychovisual criteria and can be used to calibrate the data in order to fill a larger range.

The results for the seven views are quite consistent with the average, especially for noise C_n. However, some differences related to the content exist. For example, the shadow criterion C_s has a different shape from one view to the other. The same can be said for the overall quality Q. Figure 7e and 7f show that even if two views come from the same scene, the judgment of the observers can be singularly different. Indeed, there is important variation in terms of structures, lighting effects, *etc.*

An analysis of covariance (ANCOVA) was conducted to compare one criterion in two or more groups, taking into account variability of other criteria. Table 1 gives the output of the covariance analysis for the five criteria $\{C_b, C_s, C_n, C_a, C_c\}$. For each criterion, two different questions were addressed: (1) *Is the overall quality Q related significantly to the criterion?* and (2) *For the same MOS of a criterion, does the overall quality Q vary according to the view?*

Table 1: Analysis of covariance between criteria $\{C_b, C_s, C_n, C_a, C_c\}$ and overall quality Q. For each criterion, the first line answers question (1) and the second line question (2).

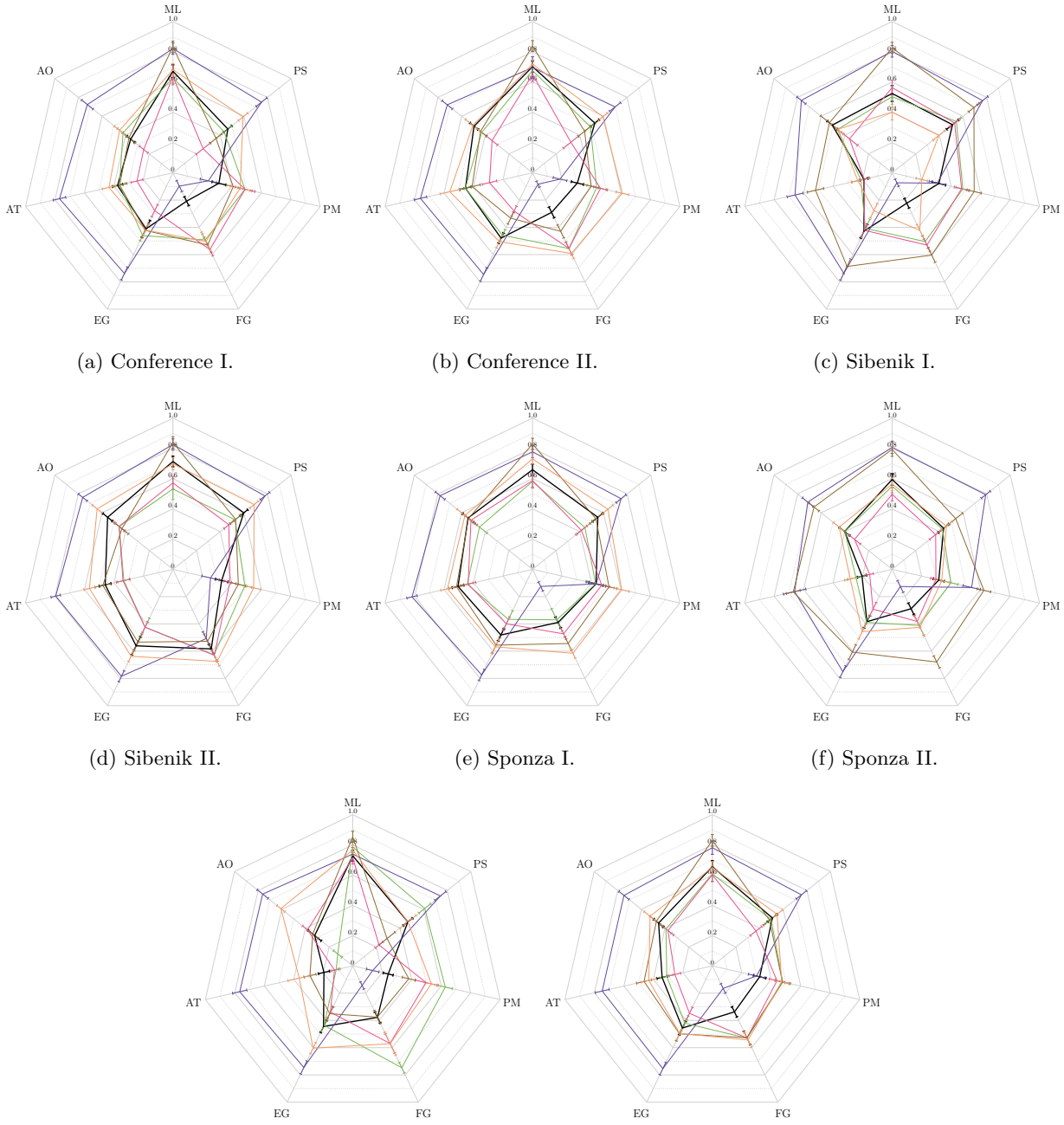
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
C_b	1	16.377	16.377	319.261	$< 2.2 \times 10^{-16}$
C_b:view	6	0.402	0.067	1.305	0.252
C_s	1	22.127	22.127	459.857	$< 2.2 \times 10^{-16}$
C_s:view	6	0.421	0.070	1.458	0.189
C_n	1	10.957	10.957	201.1214	< 2.2e-16
C_n:view	6	1.229	0.205	3.761	0.001
C_a	1	9.354	9.354	173.704	$< 2.2 \times 10^{-16}$
C_a:view	6	1.524	0.254	4.717	9.146×10^{-05}
C_c	1	32.330	32.330	778.832	$< 2.2 \times 10^{-16}$
C_c:view	6	0.205	0.034	0.824	0.551

To answer those questions, the F of *Fisher* and its associated probability (Pr) were used. Very low values of Pr mean that the variables provide significant information to the model. For example, the Pr related to contrast C_c is extremely low (compared to the threshold of 0.05), leading to the conclusion that this criterion is significantly related to the overall quality Q; it is indeed the case for the five selected criteria. For the second question, the statistical analysis indicates that appart from noise C_n and aliasing C_a, the relationship between a criterion and the overall quality Q is not affected by scene or viewpoint.

Table 2: Regression coefficients and output.

β_b	β_s	β_n	β_a	β_c	β_0	R^2	Std. Err.	DF
0.0693	0.2507	0.2138	0.0322	0.3888	-0.0519	0.507	0.1806	1375

Table 2 gives the linear regression coefficients (*cf.* Equation 5) for the five criteria. An immediate and major observation is that color bleeding C_b and aliasing C_a have significantly lower contributions (*i.e.* observers have not given a high importance to these attributes) than the other criteria. A possible explanation for this lower contribution could be the implicit combination of these attributes with the others. On the other hand, contrast C_c has an important effect on the overall quality Q (approx. 39%), followed by shadow C_s and noise C_n. In



(a) Conference I. (b) Conference II. (c) Sibenik I.

(d) Sibenik II. (e) Sponza I. (f) Sponza II.

(g) Cornell Box. (h) Average.

Figure 7: MOS and confidence intervals for the five criteria (*cf.* Section 3) on the seven views (*cf.* Figure 6) and for the seven algorithms (*cf.* Section 4.1.2). Average of the seven views is plotted in 7h.

addition to this, R^2, which measures the quality of the regression, indicates that the criteria selected explain up to 50% of the variation of the overall quality Q. It also means that there are other criteria that have to be taken into account, some of them may be implicit or difficult to formulate.

6. CONCLUSION

Study of the perceptual quality of computer-generated images through psychovisual experiments is vibrant. In this paper, such an experiment was presented, to assess the influence of five measurable psychovisual criteria, color bleeding, shadows, noise, aliasing and contrast, on the overall quality. Seven computer graphics rendering algorithms were used for the construction of the evaluation. A reprensentative panel of thirty observers participated to the experiment. The statistical analysis showed that among the five criteria, contrast, noise and shadows have a major effect on the overall quality, rather than color bleeding and aliasing. The future direction of this work is to study the correlation between automatically measured human judgment criteria.

AKNOWLEDGEMENTS

The authors wish to thank the projects UE-FEDER NAVII and ANR QUIAVU for funding, Romuald Perrot for his help on the implementation of the ML algorithm, and the thirty participants to the experiment.

REFERENCES

1. C. M. Goral, K. E. Torrance, D. P. Greenberg, and B. Battaile, "Modeling the interaction of light between diffuse surfaces," in *Proceedings of SIGGRAPH'84*, pp. 213–222, 1984.
2. G. J. Ward and P. S. Heckbert, "Irradiance gradients," in *Proceedings of Eurographics'92 Workshop on Rendering*, pp. 85–98, 1992.
3. H. W. Jensen, "Global illumination using photon maps," in *Proceedings of Eurographics'96 Workshop on Rendering*, pp. 21–30, 1996.
4. E. Horvitz and J. Lengyel, "Perception, attention, and resources: A decision-theoretic approach to graphics rendering," in *1997, Proceedings of UAI-97*, pp. 238–249, 1997.
5. W. A. Stokes, J. A. Ferwerda, B. Walter, and D. P. Greenberg, "Perceptual illumination components: a new approach to efficient, high quality global illumination rendering," in *Proceedings of SIGGRAPH'04*, pp. 742–749, 2004.
6. K. Debattista, V. Sundstedt, L. P. Santos, and A. Chalmers, "Selective component-based rendering," in *Proceedings of GRAPHITE'05*, pp. 13–22, 2005.
7. "ITU-T recommendation P.862 : An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (PESQ)," Feb. 2001.
8. "ITU-R recommendation BS.1387 : Method for objective measurements of perceived audio quality (PEAQ)," Nov. 2001.
9. "ITU-R recommendation BT.500-12 : Methodology for the subjective assessment of the quality of television pictures (PEVQ)," Aug. 2009.
10. J. T. Kajiya, "The rendering equation," *SIGGRAPH Comput. Graph.* **20**(4), pp. 143–150, 1986.
11. R. L. Cook, "Stochastic sampling in computer graphics," *ACM Trans. Graph.* **5**, pp. 51–72, 1986.
12. E. P. Lafortune and Y. D. Willems, "Bi-directional path tracing," in *Proceedings of Compugraphics'93*, pp. 145–153, 1993.
13. R. Veach and L. J. Guibas, "Metropolis light transport," in *SIGGRAPH Comput. Graph.*, **31**, pp. 65–76, 1997.
14. C. Kelemen, L. Szirmay-Kalos, G. Antal, and F. Csonka, "A simple and robust mutation strategy for the metropolis light transport algorithm," in *Comput. Graph. Forum*, pp. 531–540, 2002.
15. G. J. Ward, F. M. Rubinstein, and R. D. Clear, "A ray tracing solution for diffuse interreflection," *SIGGRAPH Comput. Graph.* **22**, pp. 85–92, 1988.
16. J. Křivánek, P. Gautron, G. Ward, H. W. Jensen, P. H. Christensen, and E. Tabellion, "Practical global illumination with irradiance caching," in *SIGGRAPH'08 classes*, pp. 60:1–60:20, 2008.

17. J. Křivánek, K. Bouatouch, S. N. Pattanaik, and J. Žára, "Making radiance and irradiance caching practical: Adaptive caching and neighbor clamping," in *Proceedings of EGSR'06 Rendering Techniques*, 2006.
18. B. Airieau, F. Bridault, D. Meneveaux, and P. Blasi, "Photon Streaming for Interactive Global Illumination in Dynamic Scenes," *The Visual Computer* **27**(3), pp. 229–240, 2011.
19. A. Keller, "Instant radiosity," in *Proceedings of SIGGRAPH '97*, pp. 49–56, 1997.
20. H. Landis, "Production-ready global illumination," in *SIGGRAPH '02 Course Notes*, **16**, 2002.
21. B. Airieau, "Light simulation with environment maps (in french)," Master's thesis, Université de Poitiers, 2008.
22. L. C. Wanger, J. A. Ferwerda, and D. P. Greenberg, "Perceiving spatial relationships in computer-generated images," *IEEE Comput. Graph. Appl.* **12**, pp. 44–51, 54–58, May 1992.
23. L. Wanger, "The effect of shadow quality on the perception of spatial relationships in computer generated imagery," in *Proceedings of I3D '92*, pp. 39–42, 1992.
24. I. Yu, A. Cox, M. H. Kim, T. Ritschel, T. Grosch, C. Dachsbacher, and J. Kautz, "Perceptual influence of approximate visibility in indirect illumination," *ACM Trans. Appl. Percept.* **6**, pp. 1–14, 2009.
25. P. Boulenguez, B. Airieau, M. Larabi, and D. Meneveaux, "Supplemental material (*e.g.* high res. CGIs and plots)." `www.sic.sp2mi.univ-poitiers.fr/spie_ist/`, 2012.
26. M. Bach, "The Freiburg visual acuity test-variability unchanged by post-hoc re-analysis," *Graefe's Archive for Clinical and Experimental Ophthalmology* **245**, pp. 965–971, 2006.