A comparison of perceptually-based metrics for objective evaluation of geometry processing

Guillaume Lavoué, Massimiliano Corsini

Abstract—Recent advances in 3D graphics technologies have led to an increasing use of processing techniques on 3D meshes, such as filtering, compression, watermarking, simplification, deformation and so forth. Since these processes may modify the visual appearance of the 3D objects, several metrics have been introduced to properly drive or evaluate them, from classic geometric ones such as Hausdorff distance, to more complex perceptually-based measures. This paper presents a survey on existing perceptually-based metrics for visual impairment of 3D objects and provides an extensive comparison between them. In particular, different scenarios which correspond to different perceptual and cognitive mechanisms are analyzed. The objective is twofold: (1) catching the behavior of existing measures to help Perception researchers for designing new 3D metrics and (2) providing a comparison between them to inform and help Computer Graphics researchers for choosing the most accurate tool for the design and the evaluation of their mesh processing algorithms.

Index Terms—Perceptual metrics, Geometry Processing, Objective Evaluation, Quality Evaluation

1 Introduction

Scientific and technological advances in the fields of telecommunications, 3D acquisition, rendering and geometry processing have boosted the diffusion of three-dimensional (3D) digital data. Nowadays, the processing, transmission and visualization of 3D objects are a part of possible and realistic functionalities over the Internet. In this context, many processing operations are commonly applied on 3D models (mostly represented by polygonal meshes) including filtering, denoising, simplification, watermarking or compression.

These operations introduce slight modifications on the 3D shape of the object, which modify its visual appearance; figure 1 illustrates some examples of processing: watermarking (method from Cho et al. [1]), simplification (QEM algorithm [2]) and denoising (Anisotropic Mean Curvature Flow [3]). The objectives of these algorithms are different, however the way they modify the visual appearance of the mesh is a critical issue for all of them. Indeed a watermarking scheme tries to maximize the size or the robustness of the mark while keeping the geometric modification as imperceptible as possible; similarly, a compression, or a simplification algorithm, attempts to minimize the stream size or the triangle number while keeping the visual difference with the original mesh as low as possible. Finally, the objective of filtering or denoising algorithms is to improve the quality of the model, while preserving as much as possible its original shape, both at the coarse and at the fine level of details.

The main problem is that these algorithms are mainly driven by geometric distances (e.g. Euclidian vertex-to-vertex distances, Hausdorff distance) which fail to correctly capture the visual quality or the perceived visual appearance difference between two 3D models. One obvious way of capturing the perceived visual impairment is to conduct subjective experiments where human observers directly give their opinion or some ratings about the processed models. However, such subjective evaluations are not only time-consuming and expensive but they also cannot be incorporated into automatic systems. In this context, several perceptually-based metrics have been proposed in the Computer

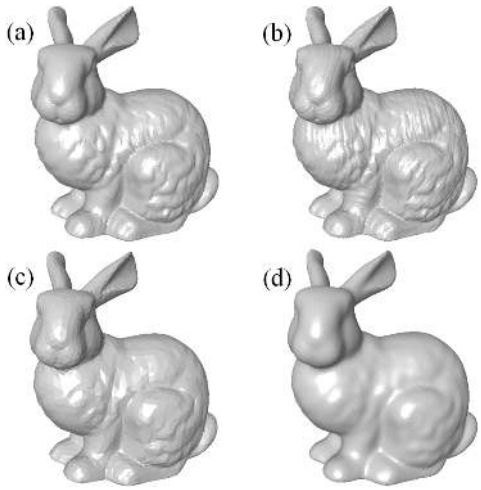

Fig. 1. Some examples of common processing operations for the Stanford Bunny mesh. *(a)* Original model (35K vertices), *(b)* result after watermarking, *(c)* result after simplification (from 35K vertices to 3.5K vertices), *(d)* result after denoising (Anisotropic Mean Curvature Flow, 7 iterations).

Guillaume Lavoué is with the Université of Lyon, CNRS, and the INSA-Lyon, LIRIS UMR 5205, FRANCE e-mail: glavoue@liris.cnrs.fr
Massimiliano Corsini is with the Visual Computing Lab, ISTI-CNR, ITALY e-mail: massimiliano.corsini@isti.cnr.it

Graphics community that aim at correctly reflecting the loss of quality or the visual difference between 3D objects. This paper gives a survey on these existing metrics together with extensive comparisons within different scenarios, to study and compare their behaviors regarding different perceptual and cognitive mechanisms. Our objective is twofold: (1) to provide an exhaustive survey and behavior understanding of the existing measures to help researchers for designing new 3D metrics and (2) to provide a comparison in some specific contexts that can help Computer Graphics researchers for choosing the accurate tool for designing or evaluating a mesh processing algorithm; that is the reason why, for several metrics, we provide also the website where it can be downloaded (see section 2.3).
The next section reviews the existing works in Computer Graphics about perception, in particular about perceptually-based measures. Section 2.3 presents more details about the metrics which are compared in this work, while sections 3 to 5 present the three different experiments aiming at analyzing the comportment of these metrics regarding three scenarios: sensitivity to the impairment frequency, sensitivity to the *masking* effect and efficiency in a general context. Section 6 discusses the respective performances of the analyzed metrics regarding some analytical criteria. Finally section 7 presents an open discussion regarding the integration of texturing and complex materials into perceptual metrics.

2 Perceptual issues in Computer Graphics

The knowledge about the human visual system (HVS) has been often exploited in the Computer Graphics community for several purposes. For example, the limits of visual perception can be used in a sophisticated rendering system in order to reduce the computational time by rendering accurately only the most visually important parts of a 3D scene. Another example is the algorithms that attempt to simplify the geometry while preserving the overall appearance of the 3D model using perceptual criteria.

In the next paragraphs, we review some of these works by categorizing them into two groups: *image-based* and *model-based* (or *geometry-based*). In the first category the perceptual mechanisms are applied on the images generated from the 3D data while in the second group the perceptual metrics work directly on the 3D model itself making the evaluation view-independent. The study proposed here treats model-based metrics designed to evaluate the quality (in terms of perceived degradation) of a specific geometry processing algorithm. As just stated, one of our aims is to understand how the tested metrics are able to predict the visual impairment of a generic geometry processing algorithm. During the review of these existing works, a brief description of some visual perceptual mechanisms will also be given.

2.1 Image-based perceptual metrics

Basically there are two different approaches to develop perceptual metrics: *mechanistic* and *black-box*. The *mechanistic* approach takes into account the complex mathematical models of the psychophysical and physiological mechanisms of the HVS in order to develop the perceptual metric. One of the most famous example of this is the Visible Difference Predictor (*VDP*) of Daly [4], an operator that is able to calculate for each pixel the probability that a human observer is able to notice differences between two images. Hence, this is a measure of fidelity of the input images. Most of these image-based metrics rely on the same visual perceptual characteristics [5], [6]:

- *The Contrast Sensitivity Function (CSF):* This function defines the contrast at which frequency components become just visible.
- *The Channel Decomposition:* The human visual system consists of several channels selective to spatial frequency and to orientation.
- *The Masking:* This effect states that a signal can be masked by the presence of another signal with different characteristics. It was firstly discovered and studied in the ambit of audio perception. Then, visual physiologists researches found similar effect for visual perception, and called it *visual masking*. Hence, visual masking concerns how a pattern with different frequency content and/or orientation influences the perception of another one.

These attributes lead to filtering operations (according to the CSF), multi-scale filter bank decompositions, errors normalizations (masking effect) and error summation across frequency bands and space.
The *black-box* approach does not rely on how the visual system works but attempt to define a function that, given the visual stimulus as input, is able to predict how much some specific visual artifacts will be perceived by a human observer. A typical example is the work of Marziliano et al. [7] which aims at detecting and quantifying blur and ringing artifacts of JPEG compression. This approach is preferable when it is difficult to determine how to integrate the different visual stimulus involved.

In Computer Graphics both mechanistic and black-box perceptual metrics have been used in three main applications: mesh simplification, perceptually-driven rendering, and evaluation of specific processes such as compression or watermarking.

Concerning *perceptually-based mesh simplification*, Lindstrom and Turk [8] proposed to render the model being simplified from several viewpoints and use a fast image quality metric, based on a simplified version of the Sarnoff Model [9], to evaluate the perceptual impact of the simplification operation. Williams et al. [10] developed a view-dependent simplification algorithm based on a simple model of CSF that takes into account texture and lighting effects. More recently, Qu and Meyer [11]

considered the visual masking effect from the 2D texture maps to lead simplification and remeshing of textured meshes.

The objective of *perceptually-driven rendering* is to determine, according to the location of the observer, the amount of accuracy to use during the rendering, for example changing the Level Of Detail (LOD) of certain models or reduce/augment sampling density in ray-tracing rendering systems. One of the first studies of this kind was the one of Reddy [12] that analyzed the frequency content in several pre-rendered images to determine for each model the best LOD to use in a real-time rendering system. Bolin and Meyer [13] used perceptual models to optimize the sampling for ray tracing algorithms. Ramasubramanian et al. [14] proposed a rendering framework to considerably reduce the overhead of incorporating a perceptual metric into a rendering system; first, they evaluated a perceptual *threshold map* taking into account the direct illumination of the scene and then such map is used to add indirect illumination, which is usually the most computational expensive task in a global illumination rendering system. Another interesting work is the one of Dumont et al. [15] that proposed a view-dependent simplification algorithm for real time rendering, based on the worst cases imperceptible contrast and spatial frequency changes.

Another important work on image-based perceptual metrics is the one of Ferwerda et al. [16], which proposed a masking model, extending the Daly *VDP* operator, which demonstrate how surface texture can hide some visual artifacts, in particular polygonal tessellation. Recently, the perceptual evaluation has been moved to a higher level of investigation concerning visual mechanisms, for example Ramanarayanan et al. [17] proposed the new concept of *visual equivalence*: images are visually equivalent if they convey the same impressions of scene appearance. In this work the authors explore how the perception of geometry, material and illumination in a scene are affected by lighting environment changes.

2.2 Model-based perceptual metrics

The main problem of the image-based metrics in the context of Computer Graphics applications is that, *in general, the perceived degradation of still images may not be adequate to evaluate the perceived degradation of the equivalent 3D model*. This has been concluded by the subjective experiments conducted by Rogowitz and Rushmeier [18]. In their work they demonstrate that the observers evaluate differently the quality of a simplified 3D model if an animation or a set of static frames of the same animation is used in the tests for the subjective evaluation. The main reason is that the object's movement introduces changes in the perception of differences that are difficult to integrate in the perceptual metric. One of the first attempts to integrate image movement, visual attention and saliency was the work of Yee et al. [19]. Myszkowski [20] proposed an extension of the VDP for quality evaluation of computer-generated animations and apply such metrics to speed-up global illumination rendering. The application of these spatiotemporal perceptual metrics in the context of 3D models visual fidelity evaluation has never been investigated from our knowledge. This is an interesting directions for future research in objects-based perceptual metrics.

Model-based metrics are used in different contexts. One of these is to control mesh simplification algorithms, in order to reduce the number of vertices while preserving the visual appearance. Kim et al. [21] state that human vision is sensitive to curvature changes and propose a *Discrete Differential Error Metric* (DDEM). In a different way, Howlett et al. [22] lead their simplification to emphasize visually salient features, determined through an eye tracking system. Lee et al. [23] follow a similar approach but automatically extract the saliency from the input mesh by computing multiresolution curvature maps. Concerning the simplification algorithm of Williams et al. [10] previously mentioned we precise that its perceptual evaluation mechanism is based also on models' geometry due to its view-dependency. Hence, this could be considered an hybrid in our categorization.

Recently, several researchers have investigated the use of black-box perceptual metrics for the evaluation of specific artifacts. Karni and Gotsman [24], in order to evaluate properly their compression algorithm, introduce the *Geometric Laplacian* (GL), which is based on a measure of the smoothness of each vertex. Pan et al. [25] propose a metric for the quality assessment of 3D models in terms of geometric and texture resolution. Their work underlines that the perceptual contribution of image texture is, in general, more important than the model's geometry. Drelie Gelasca et al. [26] and Corsini et al. [27] propose a perceptual metric based on global *roughness* variation, to measure the quality of a watermarked mesh. They gave two definitions of *roughness*, the variance of the difference between a 3D model and its smoothed version, and the variance of the dihedral angles between adjacent faces evaluated in a multi-resolution fashion. In the ambit of quality evaluation of 3D watermarking algorithms, Lavoué et al. [28] proposed a perceptually-inspired metric called *Mesh Structural Distortion Measure (MSDM)*. Most recently, Bian et al. [29], [30] developed another geometry-based perceptual metric based on the strain energy, i.e. a measure of the energy which causes the deformation between the original and the processed mesh. This metric is not specific for a certain artifact but it has been used to evaluate watermarking, compression and filtering operations.

2.3 Details about the tested metrics

Here, we provide a short description of the metrics studied in the following sections. We have considered two classical geometric measures: the Hausdorff distance (H_d) and the Root Mean Square error (RMS). We have also included two combinations of the Root Mean

Square error with the Geometric Laplacian respectively introduced by Karni and Gotsman [24] and Sorkine et al. [31]. Finally, we have considered four of the recent model-based perceptual metrics just mentioned: the Mesh Structural Distortion Measure of Lavoué et al. [28], the two roughness-based metrics developed by Corsini and Drelie Gelasca et al. [27] and the Strain Field-based metric from Bian et al. [29], [30]. We do not consider perceptual image-based metrics in our experiments since they are not reliable to predict the perceived visual impairment on 3D models for the reasons just explained at the begin of Section 2.2.

Hausdorff Distance (H_d)

The Hausdorff Distance is defined as follows: $e(p, A)$ represents the distance from a point p in the 3D space and the three-dimensional object A:

$$e(p, A) = \min_{v_i^A \in A} d(p, v_i^A) \tag{1}$$

with d the Euclidian distance and v_i^A, the i^{th} vertex of object A. Then the *asymmetric Hausdorff distance* between two 3D objects A and B is:

$$H_a(A, B) = \max_{v_i^A \in A} e(v_i^A, B) \tag{2}$$

The *symmetric Hausdorff* distance is then defined as follows:

$$H_d(A, B) = \max\{H_a(A, B), H_a(B, A)\} \tag{3}$$

In our experiments we used the *symmetric Hausdorff* distance calculated with the Metro software tool[1] [32].

Root Mean Square error (RMS)

The Root Mean Square error is based on a correspondence between each vertex of the objects to compare, hence it is limited to the comparison between two meshes sharing the same connectivity. In formula:

$$RMS(A, B) = \left(\sum_{i=1}^{n} \left\| v_i^A - v_i^B \right\|^2 \right)^{1/2} \tag{4}$$

where n is the number of vertices of the meshes and v_i^B is the vertex of B corresponding to the vertex v_i^A of A.

Geometric Laplacian measures (GL_1 and GL_2)

The Geometric Laplacian (GL) was introduced by Karni and Gotsman [24]. It is based on a measure of smoothness of the vertices. Specifically, given a vertex v:

$$GL(v) = v - \frac{\sum_{i \in n(v)} l_i^{-1} v_i}{\sum_{i \in n(v)} l_i^{-1}} \tag{5}$$

where $n(v)$ is the set of indices of the neighbors of v, and l_i the Euclidean distance from v to v_i. $GL(v)$ represents the difference vector between v and its new position after a Laplacian smoothing step. Considering (5) Karni and Gotsman [24] have derived a visual metric GL_1 between two objects A and B defined as:

$$GL_1(A, B) = \alpha\ RMS(A, B) + \\ + (1 - \alpha) \left(\sum_{i=1}^{n} \left\| GL(v_i^A) - GL(v_i^B) \right\|^2 \right)^{1/2} \tag{6}$$

with $\alpha = 0.5$. More recently, Sorkine et al. [31] proposed a different version of GL_1 (we refer to it as GL_2 which assumes a little value of α ($\alpha = 0.15$).

Mesh Structural Distortion Measure ($MSDM$)

The Mesh Structural Distortion Measure, available online[2] was introduced by Lavoué et al. [28]; this measure follows the concept of structural similarity introduced for 2D image quality assessment by Wang et al. [33]. The local $LMSDM$ distance between two mesh local windows a and b is defined as follows:

$$LMSDM(a, b) = \\ (0.4 \times L(a, b)^3 + 0.4 \times C(a, b)^3 + 0.2 \times S(a, b)^3)^{\frac{1}{3}} \tag{7}$$

L, C and S represent respectively curvature, contrast and structure comparison functions:

$$L(a, b) = \frac{\|\mu_a - \mu_b\|}{\max(\mu_a, \mu_b)} \\ C(a, b) = \frac{\|\sigma_a - \sigma_b\|}{\max(\sigma_a, \sigma_b)} \\ S(a, b) = \frac{\|\sigma_a \sigma_b - \sigma_{ab}\|}{\sigma_a \sigma_b} \tag{8}$$

with μ_a, σ_a and σ_{ab} are respectively the mean, standard deviation and covariance of the curvature over the local windows a and b. A *local window* is defined as a connected set of vertices belonging to a sphere with a given radius; this radius is a parameter of the method, we use 0.5% of the bounding box length as recommended by the authors. The global $MSDM$ measure between two meshes A and B, is defined by a Minkowski sum of their n_w local window distances:

$$MSDM(A, B) = \left(\frac{1}{n_w} \sum_{j=1}^{n_w} LMSDM(a_j, b_j)^3 \right)^{1/3} \in [0, 1) \tag{9}$$

where n_w is the number of local windows of the meshes and b_j is the local window of B corresponding to the window a_j of A. Practically, this measure considers one local window per vertex of the original mesh and is asymmetric. Its value tends toward 1 (theoretical limit) when the measured objects are visually very different and is equal to 0 for identical ones.

1. http://vcg.isti.cnr.it/activities/surfacegrevis/simplification/metro.html
2. http://liris.cnrs.fr/guillaume.lavoue/rech/soft.html

Roughness-based Measures ($3DWPM_1$ and $3DWPM_2$)

Following the idea that a measure of the visual artifacts produced by watermarking should be based on the amount of roughness introduced on the surface, Corsini and Drelie Gelasca et al. [27] proposed two perceptual metrics for quality evaluation of watermarking algorithms; these two metrics will be available in a future release of the Meshlab software[3].

The watermarking visual impairment is evaluated by considering the increment of roughness between the original model A and the watermarked model B in the following way:

$$3DWPM(A,B) = \log\left(\frac{\rho(B)-\rho(A)}{\rho(A)} + k\right) - \log(k) \quad (10)$$

where $\rho(A)$ is the total roughness of the original model and $\rho(B)$ is the total roughness of the watermarked model. The constant k is used to avoid numerical instability. Two ways to measure model's roughness are proposed.

The first roughness measure ($3DWPM_1$) is a variant of the method by Wu et al. [34]. This metric measures the per-face roughness by making statistical considerations about the dihedral angles, i.e. the angle between the normals of two adjacent faces. The idea [35] is that the dihedral angle is related to the surface roughness. In fact, the face normals of a smooth surface vary slowly over the surface, consequently the dihedral angles between adjacent faces are close to zero. In order to take into account the *scale* of the roughness the per-face roughness is turned into a per-vertex roughness and rings of different size (1-ring, 2-ring, etc.) are considered during roughness evaluation. The total roughness of the 3D object is the sum of the roughnesses of all vertices.

The second method by Drelie Gelasca et al. [26] ($3DWPM_2$) is based on the consideration that artifacts are better perceived on smooth surfaces. Following this statement, this approach applies a smoothing algorithm and then measures the roughness of the surface as the variance of the differences between the smoothed version of the model and its original version.

Strain Field-based Measure (SF)

This is the most recent model-based perceptually-motivated metric to evaluate meshes' deformations. It is based on the *strain energy* introduced by the mesh deformation. The idea is that the higher the mesh is deformed, the higher is the probability that the observer perceives the difference between the processed and the original mesh. The strain energy calculation on the mesh is simplified considering that each mesh element (a triangular mesh is assumed) is perturbed along its plane. For the details about the simplification assumptions we refer to the original papers [29], [30]. It is important to underline that this metric is suitable for small deformations. The perceptual distance $SF(A,B)$ between the original model A and the perturbed one B is defined as the weighted average strain energy (ASE) over all triangles of the mesh, normalized by the total area of the triangular faces (S):

$$SF(A,B) = \frac{1}{S}\sum w_i W_i \quad (11)$$

w_i are weights and W_i is the strain energy associated to each triangle. Varying the w_i weights Zhe Bian et al. tested some variants of this metric, but from their experimental results they concluded that the simpler one $(w_i = 1)$ gave results similar to the other variants, hence the unweighted one is preferable due to its simplicity. In the next sections, we use our own implementation of this metric since the original implementation is not publicly available. For this reason our conclusions have to be considered qualitative and not quantitative due to small numerical differences that the different implementation could gives.

3. http://meshlab.sourceforge.net/

3 First experiment: sensitivity to the impairment frequency

It is now admitted in the Computer Graphics community that, in general, high-frequency distortions of a 3D shape have a high probability to be *visually* noticeable. This is because the human eye is more sensitive to the rapid variations of the surface (which directly influence the intensity image after rendering) than to lower frequency modifications like slight stretching of the whole shape. This phenomenon is illustrated in figure 2 where low and high frequency distortions applied on the Bimba model are shown; the top right model is associated with a higher geometric distortion than the bottom right model, however the perceptual impairment caused by the high frequency perturbations (bottom right) is clearly more visible. This perceptual mechanism has been already employed in several compression [31] and watermarking [36], [37] methods to hide the distortions introduced by the processing.

3.1 Goals and Motivations

The objective of this experiment is to examine if the studied metrics follow this principle from a qualitative viewpoint, i.e. if a certain metric is able to provide small distances for low-frequency impairments and high distances for high frequency ones. Obviously, also the amplitude of the frequency affects the final perception of the perturbation, for this reason, an in-depth analysis of frequency sensitivity would require many evaluations. This is not the case for this experiment since our evaluation is more qualitative than quantitative. More specifically, we would like to demonstrate that the geometric metrics, even the sophisticated ones like the Hausdorff distance and the others based on geometric Laplacian, are not able to catch this phenomenon well. On the

contrary, the perceptually-motivated metrics are able to model well this perceptual mechanism even if they do not rely on frequency analysis. In particular, we expect that the Hausdorff (H_d) distance could work well when the distortions are applied on medium-high frequencies due to its geometric properties, but we also expect, from its definition, that it is not able to evaluate correctly the impairment of low distortions with high amplitude. Concerning the two versions of the geometric Laplacian metrics (GL_1 and GL_2) it is quite difficult to evaluate intuitively their behaviors. Geometric Laplacian is related to the smoothness of the surface, hence in many cases it is reasonable to expect a good measurement of the impairment for low frequency distortions. The two variants of the $3DWPM$ metric have been explicitly designed to account for roughness variations at multiple scales, hence we expect good results from this kind of metric. The same reasoning could be made for the $MSDM$ perceptual metric. The strain energy-based one (SF) should also work well, but, due to its simplification assumptions that make it particularly suitable for small perturbations, the range of frequencies and amplitudes that can capture could be restricted with respect to the other perceptually-based metrics.

3.2 Corpus description

According to the objectives just described we have *modified* some models with high and low frequency distortions using a spectral analysis tool and then evaluated the metrics under investigation on such models.
Because of the irregular sampling intrinsic to a 3D mesh, the classical mathematical tools for spectral analysis, like the Fourier Transform, are not available for this kind of data. For this reason, several mathematical tools related to frequency analysis of meshes have been developed in the last years to fill this gap. A very recent one is based on the *Manifold Harmonics* basis introduced by [38] (composed of the eigenfunctions of the Laplace-Beltrami operator) that seems to approximate well a real Fourier Transform on the 3D mesh domain. Hence for two 3D models: *Bimba* and *Dyno*, we have produced low frequency modified versions and high-frequency modified versions. We have chosen these two models since they are very different: the Bimba model (see figure 2) is rather convex and contains a majority of smooth parts while the Dyno model (see figure 4) has a complex shape and is rough almost everywhere. The modified versions are produced by adding a binary uniform random noise on respectively the first and the last 100 coefficients of the frequency spectrum, and then reconstructing the models. The noise is applied according to three different strengths: 0.02%, 0.01% and 0.005% of the square bounding box length. An example is provided in figure 2; the top right object was modified on low frequencies (strength=0.02) while the bottom objects were modified on high frequencies (strength=0.01 and 0.005). The colored vertex displacement maps confirm the frequency of the distortions.

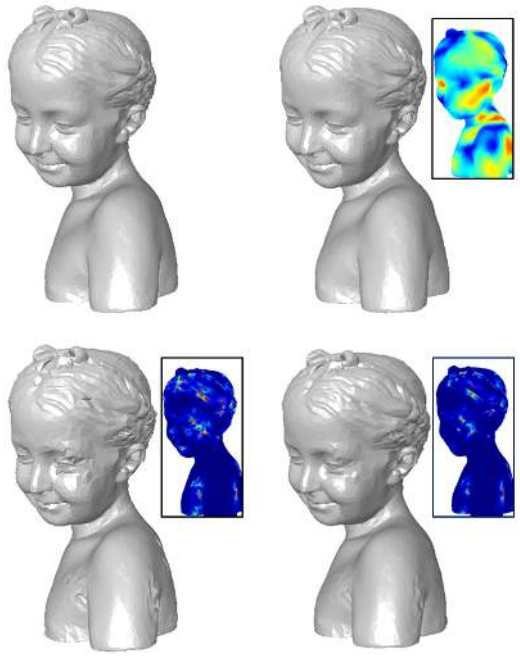

Fig. 2. Visual effect of the impairment frequency. *Top:* Original Bimba model (9K vertices) (left) and result after random noise addition on low frequencies (strength=0.02) (right). *Bottom:* Results after random noise addition on high frequencies, with strength=0.01 (left) and strength=0.005 (right). The distance maps regarding the original model are also provided (warmer colors represent higher values).

3.3 Results and discussion

Table 1 presents the results of the different metrics regarding Bimba and Dyno modified models. Figure 2 shows clearly that the impairments produced by the high frequency-*low* amplitude noise (bottom/right in the figure, $HF_{0.005}$ in table 1) are much more visible than those produced by the low frequency-*high* amplitude noise (top/right in the figure, $LF_{0.02}$ in table 1); here we analyze how the tested metrics follow this perceptual mechanism.

Before we discuss the results obtained, we recall that some of them are normalized into a given range of values while others are not normalized. In particular, RMS, GL_1, GL_2 and H_d are not normalized, $MSDM$ is normalized into the $[0,1]$ range, the two $3DWPM$ metrics follow the score of the subject in the $0-10$ range and the SF metric is not normalized into a range of values but it is subdivided for the model's surface area.

Firstly, we point out that the behaviors of the different metrics are perfectly consistent for both models Bimba and Dyno. As expected, the Root Mean Square provides poor results ($RMS(LF_{0.02}) \gg RMS(HF_{0.005})$), however the relevance of combining RMS with the geometric Laplacian measure (which reflects the smoothness of the surface) is clearly demonstrated by this experiment since GL_2 provides better results than GL_1 which provides

TABLE 1
Metrics' values for the different high (HF) and low (LF) frequency distortions applied on the Bimba and Dyno models.

	RMS	GL_1	GL_2	H_d	$MSDM$	$3DWPM_1$	$3DWPM_2$	$SF(10^{-3})$
Bimba								
$LF_{0.02}$	0.35	0.18	0.06	5.6	0.19	2.5	2.0	1.34
$LF_{0.01}$	0.17	0.09	0.03	2.9	0.11	1.3	1.4	0.34
$HF_{0.01}$	0.17	0.17	0.18	8.2	0.46	6.5	6.1	21.18
$HF_{0.005}$	0.09	0.09	0.09	4.1	0.35	5.0	4.4	5.3
Dyno								
$LF_{0.02}$	0.35	0.18	0.06	3.7	0.10	1.7	1.1	1.74
$LF_{0.01}$	0.17	0.09	0.03	1.8	0.07	0.5	0.3	0.16
$HF_{0.01}$	0.17	0.18	0.18	7.3	0.36	6.7	6.2	23.51
$HF_{0.005}$	0.09	0.09	0.08	3.6	0.28	5.3	4.7	5.88

better results than RMS, in particular $GL_2(LF_{0.02}) < GL_2(HF_{0.005})$. It is interesting to notice that the Hausdorff distance is able to follow human perception in some cases; the main reason is that the high frequency noise can cause severe vertex displacements (kinds of sharp bumps on the surface) which increase the H_d distance more than a low frequency deformations. However, for very small distortions on high-frequencies or very large distortions on low frequency, the metric fails, indeed for our examples $H_d(LF_{0.02}) > H_d(HF_{0.005})$. All the perceptually-based metrics ($MSDM$, $3DWPM$ and SF) have a very good behavior regarding this phenomenon, their values for $LF_{0.02}$ are clearly below those for $HF_{0.005}$. We can conclude that, even if they are not directly based on frequency analysis, these metrics are much less sensitive to low frequency noise than to high frequency noise for a certain range of amplitude distortions, just like the human visual system.

4 Second experiment: sensitivity to the Masking Effect

The objective of this experiment is to evaluate the behavior of the examined metrics regarding the visual masking effect. As explained previously, this perceptual mechanism regards how a visual pattern with specific characteristics of orientation and frequency content is hidden by another pattern with different characteristics. In our context this concept can be remapped as the fact that the human eye cannot distinguish a small distortion if it is located on a *rough* (or *noisy*) area. This property is particularly interesting for compression or watermarking to concentrate the compression artifacts or the watermark strength on *rough* parts where geometric modifications are nearly invisible, as in [39]. To establish the efficiency of the metrics regarding this phenomenon, we have considered a corpus where some noise has been added either on smooth or rough parts of several 3D models. Each distorted model of this corpus is also associated with subjective Mean Opinion Scores (MOS) from human observers. MOS reflects the observers' opinions regarding the visual difference between the original and the processed shape. The objective is to study for the corpus objects the correlation between the observers' rates and the metrics' values. The corpus and the MOS have been collected from [39]; nevertheless some details about the corpus construction and the subjective evaluation protocol are given in the following sections, since these details are relevant for the present work.

4.1 Corpus description

To construct this experimental corpus, different 3D models were considered: *Armadillo*, *Dyno*, *Lion Head* and *Bimba*. These models have been selected since they all contain both smooth and rough parts. The *roughness* was calculated for each vertex of these meshes using the estimator from [39], and then classified (K-means algorithm) into two clusters: rather rough vertices and rather smooth vertices. Figure 3 illustrates the roughness map and the two clusters obtained for the Lion Head model.

A uniform random noise was then applied only on vertices from smooth and rough clusters respectively; this noise was applied on rough and smooth regions with different strengths such as to obtain the same RMS error in each case. In other words, if the total area of the rough region was larger than the area of the smooth one, the noise applied on the rough region was lower (and vice versa). These noise distortions were applied according to three strengths (visually chosen): high, medium and low. Hence this experimental corpus contains 28 models (4 originals + 4×3 versions with noise on smooth parts + 4×3 versions with noise on rough parts). Figure 3 illustrates two models from this corpus: the last two models on the right have respectively noise on rough and smooth parts (medium distortion); for this example the noise strength is slightly higher for smooth regions: 0.155% against 0.150% of the length of the cubic bounding box of the model. Both noisy versions are associated with the same RMS distance from the original model ($1.04 \cdot 10^{-3}$). As expected the visual distortion is far less visible for the object on the left thanks to the masking effect.

4.2 Subjective evaluation protocol

The evaluation protocol is as follows: first, in the training phase, some original and distorted models from the

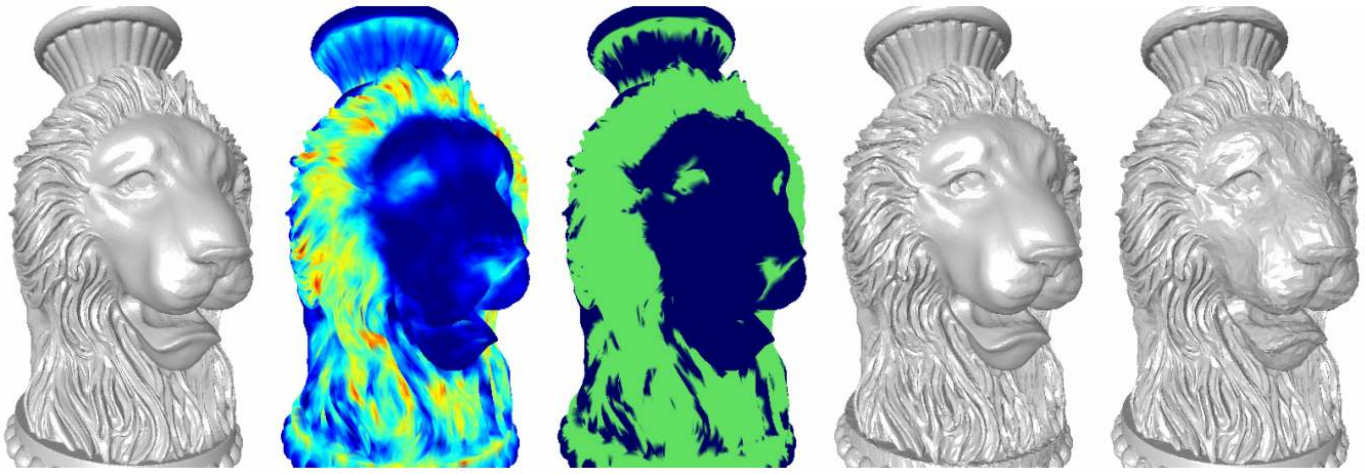

Fig. 3. Construction and example of the Masking Effect corpus. *From left to right:* Original Lion Head model (39K vertices); roughness values (warmer colors represent higher values); result of the clustering: rough vertices are in green and smooth ones are in blue; random noise on rough regions; random noise on smooth regions.

corpus were displayed to the observer so that he could get used to the shape and the strengths of the distortions. Then for each of the four models from the corpus (Armadillo, Dyno, Lion Head and Bimba), the corresponding 6 degraded versions were displayed to the observer together with the original object. Then the observer was asked to provide a score, for each object, reflecting the degree of perceived visual similarity, between 4 (identical to the original) and 0 (worst case). The objects were displayed during about 3 minutes and user interaction was allowed (rotation, scaling, translation). Following the considerations of [18] the interaction should improve the reliability of the subjective experiment in this context (this approach has just been used with positive results in [35], [26] and successive studies). It is important to note that since the observer can see all the 6 degraded versions on the same screen, there was no need to establish a referential range, since he naturally puts a 0 for the object he finds the most degraded and 4 to the best one. In order to avoid the effect of the spatial and temporal sequencing factors, the sequence of 4 models and 6 degraded versions were randomly shown for each human observer. The mean opinion score (MOS) is then computed for each noisy object of the corpus:

$$MOS_i = \frac{1}{n}\sum_{j=1}^{n} m_{ij} \tag{12}$$

where MOS_i is the mean opinion score of the i^{th} object, n is the number of test subjects, and m_{ij} is the score (in the range $[0, 4]$) given by the j^{th} subject to the i^{th} object. This subjective experiment has been conducted on 11 researchers (students and staff) from the Université de Lyon (France).

4.3 Statistical analysis

Before exploiting the results of the mean opinion scores, we have to analyze and process them; first, since some observers may have not used the whole available rating range, a subject-to-subject correction is done by normalizing the gain and offset among the observers, similarly to [40]. Second, a screening of possible outlier subjects is performed according to recommendation of the I.T.U. (International Telecommunication Union) [41]. One outlier was detected out of the 11 subjects.
Then, in order to check the suitability of our evaluation protocol and the relevance of the mean opinion scores, we assessed the variation between the different observers in their subjective ratings of the objects. The value of the intraclass correlation coefficient [42] (ICC) is 0.76, that is a very good value that means that the observers had a very good agreement on their visual estimations; hence we can assert that our protocol was correct since it led to produce meaningful consistent ratings among the observers.

To study the correlation between the mean opinion scores and the metrics' values we have considered two statistical measures [43]:

- The Pearson Product Moment Correlation (Pearson's correlation for short), that is a measure of the *linear* dependence between two variables. It is obtained by dividing the covariance of the two variables by the product of their standard deviations.
- The Spearman's rank correlation, that is a *non-parametric* measure which is based only on the ranks of each variable, not on their values. It measures the monotonic association between the variables, no assumption are made on their relationship. Basically it is calculated in the same way that Pearson's correlation, after having replaced the variables by their rank-orders.

Before computing these values, we operate a *psychometric curve fitting*, to optimize the matching between the values given by the objective metrics and the subjective scores

provided by the subjects. This step allows to take into account the saturation effects typical of human senses. For these reasons, psychometric curves exhibit a typical sigmoid shape that penalizes the strongest stimuli. This fitting allows us to evaluate the performance of the perceptually-based measures, but could also be included in the perceptual metrics for specific applications like in [27], [28]. We used in this work, the Gaussian psychometric function (recommended by [27]), which has been applied to every tested metrics:

$$\mathcal{M}_{\text{fit}} = g(a, b, M) = \frac{1}{2\pi} \int_{a+b\mathcal{M}}^{\infty} e^{-\frac{t^2}{2}} dt \qquad (13)$$

where $\mathcal{M}$ is the perceptually-based metric used, a and b are the parameters estimated by a nonlinear least-squares data fitting.

4.4 Correlation results and discussion

Table 2 presents Spearman and Pearson correlations for the different metrics (after psychometric curve fitting) and the MOS for the noisy models of the corpus. These values are given per model and also for the whole corpus. However, we have to be precise that the correlations over the whole set of models (Whole Corpus) are not really meaningful since the referential range for the rating was established separately for each model. Hence, also in this experiment our final considerations have to be considered more qualitative than quantitative. The means and the standard deviations, over the models, are also provided in the table.

One first observation is that the results are not good, except for the $MSDM$. This happens since the masking phenomenon is a complex effect and almost all these metrics do not try to model it explicitly. Both the 3DWPM metrics, which are based on the global *roughness* of the surfaces, fail to capture this cognitive phenomenon; the main reason is that the masking effect is a very *local* phenomenon, whereas these measures are based on *global* differences. It is interesting to notice that, contrary to the previous experiment, the simple RMS distance provides better results that the geometric Laplacian-based ones; its Spearman correlation is high for each object (Mean=70%), moreover its results are very stable ($\sigma = 2.9\%$); however its Pearson values, which define the real strength of the relationship, are weak (Mean=35.1%). Concerning the $MSDM$ measure it outperforms the other metrics on this corpus; these good results can be explained by two reasons: first it is based on *local* measures over the surfaces and second it relies on curvature statistics which are strongly linked with the *roughness* of the surface (and hence with masking). In particular, for the Lion model which exhibits a very high masking region (the *mane*), the $MSDM$ measure has a very good Pearson correlation (78%), while other metrics lead to particularly low values ($< 25\%$). Even over the whole set of models, the Spearman correlation of this metric is quite good (65.2%). A final considerations regards the Bimba model which is associated with low correlation values for all the metrics; the reason is probably that it represents a human face. Human face images are well-known in subjective experiments as a high-level factor attracting human attention, i.e., the distortions are perceived differently by the human observers, often as more visible, on this kind of model with respect to other ones.

5 Third experiment: behavior in a general-purpose context

In this section, our ideal goal is to evaluate the different perceptually-based metrics in a general-purpose context, in order to establish their efficiency in any kind of mesh processing scenario. In practice, we consider a specific set of mesh processing operations (noise/smoothing) trying to cover most of the possible distortions/modifications occurring in common geometry processing algorithms. We leave a more complete study of this kind as a matter of future research. In the following sections we give the details about the corpus construction (an extension of the one used in [28]) and about the subjective evaluation protocol used. We also discuss the difficulties to make an experiment of this kind more general.

5.1 Corpus description

We have considered four models, widely used in the Computer Graphics community: *Armadillo*, *Dyno*, *Venus* and *RockerArm*; these models have been chosen for their very different characteristics: Armadillo and Dyno present complex shapes, with many rough areas, Venus is rather convex with a majority of smooth parts and RockerArm is completely smooth and of genus one. Moreover these models represent several different applications: Computer-Aided Design (RockerArm), Cultural Heritage (Venus) and Video Games (Armadillo and Dyno).

We have then applied two types of distortions on these models: noise addition and smoothing (achieved with the technique of Taubin et al. [44]). These distortions were applied according to three strengths (visually chosen): high, medium and low (these strengths correspond to a number of iterations for smoothing and a value of maximum deviation for noise addition). Moreover, these distortions where applied also on different locations: uniformly (on the whole object), on rather *smooth* areas, on rather *rough* areas and on *intermediate* areas. The roughness was simply defined, in that case, by the variance of the curvature. Since the smoothing was not applied on smooth areas, 21 degraded versions were produced per model (3 noise strengths $\times$ 4 locations + 3 smoothing strengths $\times$ 3 locations), and thus the experimental corpus contains 88 models (4 originals + 4$\times$21 degraded versions). Figure 4 presents some samples of the corpus. These non-uniform noise addition and smoothing basically reflect a lot of possible distortions

TABLE 2
Spearman (r_S) and Pearson (r_P) correlation values (%) between Mean Opinion Scores and values from perceptually-based metrics for the masking corpus. Mean values and Standard deviations over all models are also given.

	RMS		GL_1		GL_2		H_d		$MSDM$		$3DWPM_1$		$3DWPM_2$		SF	
	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P
Armadillo	65.7	44.6	65.7	44.4	65.7	44.2	48.6	37.7	88.6	72.2	58.0	41.8	48.6	37.9	48.6	40.4
Lion	71.4	23.8	37.1	22.4	20.0	21.6	71.4	25.1	94.3	78.0	20.0	9.7	38.3	22.0	20.0	1.8
Bimba	71.4	21.8	20.0	19.8	20.0	18.0	25.7	7.5	42.9	33.9	20.0	8.4	37.1	14.4	20.0	12.3
Dyno	71.4	50.3	71.4	50.0	60.0	49.8	48.6	31.1	100.0	91.7	66.7	45.3	71.4	50.1	88.5	54.1
Whole Corpus	48.8	17.0	42.0	15.7	40.1	14.7	26.6	4.1	65.2	47.9	29.4	10.2	37.4	18.2	38.6	2.4
$Mean$	70.0	35.1	48.6	34.1	41.4	33.3	48.6	25.3	81.4	69.0	41.2	26.2	48.8	31.1	45.3	35.8
σ	2.9	14.5	24.2	15.3	24.9	16.0	18.7	12.9	26.1	24.8	24.7	20.0	15.9	16.0	34.4	38.8

occurring during common geometric processing operations such as:

- Denoising filters; the final results of such filters is often similar to smoothing.
- Compression; many algorithms are based on a geometric quantization which introduces noise-like deformations.
- Watermarking; many algorithms introduce structured noise-like deformations.

It could probably have been better for our purpose to create a corpus containing meshes processed using several existing compression, denoising/filtering or watermarking algorithms; however creating such a corpus is a challenging task by itself because the different distortions have to stay in the same visual perceptual range to produce relevant results and that constraint is difficult to resolve for distortions of very different natures, in particular if mesh simplification algorithms are also taken into account. For this reason we attempt to simulate the visual impairment of generic geometric processing operations (except simplification and remeshing) with noise/smoothing operations. Another problem in the construction of a corpus more general than this is that a lot of 3D models have to be considered and a subjective experiment which takes a long time to be done could make the observer's scores less reliable.

5.2 Subjective evaluation protocol

The evaluation protocol of this experiment basically follows the one defined by Corsini et al. [27]. First, the original models were displayed together with some distorted ones and with the worst cases (uniform, maximum strength) for noise and smoothing in order to establish a referential range for the rating (this constitutes the training phase). It is important to underline that for all the objects, both worst cases (noise and smoothing) where displayed and the subject was asked to remember the one he found the worst among them. Finally, the 88 objects of the corpus were displayed one by one on the screen, each for 20 seconds, and the subjects were asked to provide a score reflecting the degree of perceived distortion, between 0 (identical to the original) and 10 (worst case). In order to avoid the effect of the temporal sequencing factor, the presentation sequence of the 88 objects was randomly generated for each participant. Like in the previous experiment, user interaction was allowed (rotation, scaling, translation). This subjective experiment has been carried out on a pool of 12 students from the Swiss Federal Institute of Technology (Lausanne, Switzerland) and from the Université de Lyon (France).

5.3 Statistical analysis

Like in the previous experiment, before exploiting the results we have conducted normalization and screening of the MOS; one outlier was detected out of the 12 subjects. The value of the Intraclass Correlation Coefficient (ICC), which measures the agreement of the observers on their ratings, is 0.60; this value is not as high as for the previous experiment but remains quite good. Hence, we can assert that our protocol was reliable since ratings are quite consistent among the observers.

5.4 Correlation results and discussion

The Spearman and Pearson correlations between the collected MOS and the values of the metrics are reported in Table 3. For each object three correlation values are presented: *All distortions*, *Smoothing* and *Noise*; their calculations are respectively based on the 21 distorted (smoothing + noise) versions, the 9 smoothed versions and the 12 noisy versions; the original model is also taken into account. The correlations over the whole corpus are also given for each metric; these correlation values are meaningful in this experiment (contrarily to the previous one) because in this subjective evaluation protocol we establish one single referential range for all the set of models. Moreover, these correlation values are very useful since they illustrate the capacity of the metrics to correctly compare visual impairments from different models.

The first point to make is that, when considering only one type of modification (noise or smoothing) and only

Fig. 4. Samples of the general-purpose corpus. *From left to right:* Venus with noise on *rough* areas, Venus with global smoothing, Armadillo with noise on *smooth* areas, RockerArm with noise on *intermediate* areas, Dyno with smoothing on *rough* areas.

TABLE 3
Spearman (r_S) and Pearson (r_P) correlation values (%) between Mean Opinion Scores and values from perceptually-based metrics for the general-purpose corpus. Mean values and Standard deviations per models are also given.

	RMS		GL_1		GL_2		H_d		$MSDM$		$3DWPM_1$		$3DWPM_2$		SF	
	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P	r_S	r_P
Armadillo																
Smoothing	69.7	50.7	64.9	46.7	64.9	41.9	70.3	29.0	78.2	51.6	66.9	30.7	64.4	30.3	64.9	52.8
Noise	89.6	84.4	89.6	84.5	89.6	84.6	76.8	48.8	84.1	72.8	87.2	74.0	89.0	73.6	89.6	91.1
All distortionss	62.7	32.2	70.2	43.7	77.8	55.5	69.5	30.2	84.8	70.0	65.8	35.7	74.1	43.1	51.2	16.9
Dyno																
Smoothing	72.1	26.9	72.1	22.4	70.9	17.2	56.3	18.2	67.3	24.1	26.1	6.1	48.0	11.0	57.6	15.5
Noise	93.4	86.8	93.4	86.7	93.4	86.7	83.4	73.9	90.1	79.9	86.8	58.7	90.9	59.0	93.4	86.3
All distortions	0.3	0.0	15.5	3.2	30.6	12.5	30.9	22.6	73.0	56.8	62.7	35.7	52.4	19.9	0.4	0.7
Venus																
Smoothing	89.1	68.8	91.5	65.9	95.2	71.2	79.9	51.2	86.7	62.3	70.5	41.9	87.9	70.5	92.7	89.4
Noise	89.6	87.3	89.6	87.2	89.6	87.1	76.9	61.2	85.2	77.4	84.1	67.8	80.2	58.4	89.6	87.2
All distortions	90.1	77.3	92.0	80.2	91.0	77.6	1.6	0.8	87.6	72.3	71.6	46.6	34.8	16.4	87.4	69.0
Rocker																
Smoothing	83.0	75.3	83.0	73.4	83.0	70.8	63.0	40.7	91.5	80.9	76.6	35.5	75.4	44.1	83.0	71.9
Noise	95.6	94.7	97.3	97.7	97.3	94.8	84.6	84.2	75.8	70.6	97.7	95.6	96.8	95.6	97.3	82.6
All distortions	7.3	3.0	14.2	8.4	29.0	17.1	18.1	5.5	89.8	75.0	87.5	53.2	37.8	29.9	6.8	0.5
Whole Corpus																
Smoothing	54.5	25.8	51.7	23.1	48.2	20.2	45.8	27.1	74.9	54.1	57.4	27.6	66.0	31.5	45.6	12.7
Noise	68.7	47.9	68.3	47.5	68.1	47.0	47.1	23.8	72.9	56.6	72.2	47.0	87.7	69.6	68.4	38.4
All distortions	**26.8**	**7.9**	**33.1**	**12.6**	**39.3**	**18.0**	**13.8**	**1.3**	**73.9**	**56.4**	**69.3**	**38.3**	**49.0**	**24.6**	**15.7**	**0.5**
$Mean$	40.0	28.1	48.0	33.9	57.1	40.4	30.0	14.7	83.8	67.6	71.9	42.5	49.8	27.2	36.4	21.8
σ	43.6	35.9	39.2	35.7	31.9	30.9	28.9	13.9	7.5	8.0	11.0	8.9	18.0	12.1	40.8	32.4

one 3D object, every metric leads to quite high correlation values whatever the object. This is caused by the psychometric fitting. Hence, the real measure of strength of a metric is its capacity to provide high correlation values over several models processed in different ways. When considering only noise distortions, the metric which provides the best results over the whole corpus is $3DWPM_2$, its Pearson correlation is quite high ($r_P = 69.6\%$) compared with its counterparts ($r_P = 56.6\%$ for MSDM, and $r_P < 50\%$ for the others). We recall that this metric relies on the roughness computed as the differences between the model and its smoothed version; this mechanism seems to be very efficient to capture the perception of noise and thus explains the good results of $3DWPM_2$. When considering only smoothing distortions, correlation values over the whole corpus are quite low for all the metrics; the effect of smoothing is not evident to the human visual system as the addition of noise. $MSDM$ provides the best results, its Pearson correlation is rather low ($r_P = 54.1\%$) but much higher than the other metrics ($r_P < 32\%$).

When considering both types of distortions (smoothing and noise), it becomes much more difficult for the metrics to correlate with human perception, even when considering one model at a time, since they have to be able to correctly merge the visual effects produced by noise and smoothing. In this difficult scenario, purely geometric RMS and Hausdorff distances completely fail (their

Sperman values for the whole corpus are respectively 26.8% and 13.8%). As for the first experiment (see section 3), the combinations of RMS with the geometric Laplacian provide better results: respectively 33.1% and 39.3% Spearman correlation values for GL_1 and GL_2. The most advanced metrics work well despite the difficulty of the context: $MSDM$ and $3DWPM_1$ lead respectively to 73.9% and 69.3% in term of Spearman correlation. Regarding Pearson correlation, $MSDM$ performs better than its counterparts (56.4% while other are $< 39\%$). Figure 5 presents the psychometric curve fitting between the objective and subjective scores for RMS, GL_2 and $3DWPM_1$ for the Dyno model; it illustrates clearly the difficulty to correctly merge the impairments produced by noise addition and smoothing (white and blue circles respectively); this figure illustrates also the improvement brought by the use of the geometric Laplacian in GL_2 over the simple RMS and the superior performances of perceptual ($3DWPM_1$, $MSDM$) vs non-perceptual (RMS, GL_2) metrics.

Figure 6 illustrates the psychometric curve fitting between the objective and subjective scores over the whole corpus. These sub-figures are very interesting since they summarize visually the performances of the tested metrics. Each model is represented by a different symbol. The respective performances of the metrics are confirmed; indeed, the fitting (i.e. the possible prediction) is more efficient for $MSDM$ and $3DWPM_1$. In particular, the prediction error (RMSE) is lower for these metrics (1.31 for $MSDM$ and 1.57 for $3DWPM_1$). Finally, we can observe that the SF metric provides poor results (15.7% for Spearman correlation over the whole corpus), the main reason is that its values regarding smoothing modifications are almost always higher than for noise modifications, while the resulting visual alterations are clearly lower. This difficulty to merge the visual effects produced by noise and smoothing is illustrated in figure 7: the metric demonstrates a very high correlation with subjective scores for smoothed and noisy versions separately, however when considering both distortions together the correlation is very poor (the smoothing distortions are overestimated). We have to notice, though, that this metric assumes the distortions to be small, hence its poor behavior can be caused by the quite high strength of the distortions in our corpus.

One critical feature of a good metric is the *stability* of its results; table 3 details the values of standard deviation (σ) of the Spearman and Pearson correlations among the 4 objects Armadillo, Dyno, Venus and RockerArm. Once again the best metrics according to this criterion are $MSDM$ and $3DWPM_1$ which exhibit a very good stability ($\sigma_{Pearson}$ are respectively 8.0% and 8.9%). All the other metrics have a lower robustness, in particular they all provide very poor results for Dyno and RockerArm.

6 Computational comparison

6.1 Algorithm requirements

The analyzed metrics need to be compared also from a computational point of view since they have different requirements to work correctly. For example, some metrics are independent on the connectivity of the meshes to compare while others require the same connectivity (i.e. the same number of vertices), some metrics are influenced by the vertices density, i.e. the level of detail of the mesh, and so on.

Hausdorff distance is certainly the most compliant metric; it can compare any kind of meshes even if the levels of detail or the connectivity of the mesh to compare are different.

Both the *3DWPM* metrics present the same (not strict) constraints. Since they are based on global roughness difference they can theoretically compare objects with different connectivities, however, since the roughness is calculated on vertex neighborhoods they are dependent on the sampling density even if the $3DWPM_1$ reduces this dependence by using a multi-resolution approach. Moreover, uniform sampling is assumed.

The $MSDM$ measure is linked to a scale parameter which makes it independent on the connectivity of the meshes to compare; hence in theory any kind of meshes, even with different levels of details could be compared. In its current implementation (available on-line), the meshes have to share the same connectivity and the same vertex order in the files, this constitutes a heavy constraint but it is an implementation issue and not an intrinsic limit of the algorithm.

Finally, for RMS, GL_1, GL_2 and SF which are based on vertex-to-vertex mapping, the meshes have to be *consistent*, i.e. they have to share the same connectivity.

6.2 Processing time

Table 4 details the processing times for different model sizes. For each value, the object has been compared with itself, on an Intel Core 2 Duo processor with 2GB memory. Of course since they operate simple vertex-to-vertex measures, RMS, GL_1 and GL_2 are particularly fast; SF is also very fast, for the same reason. Hausdorff distance, $MSDM$ and $3DWPM_2$ present similar processing times (around 15-20 seconds for the Feline model), with similar linear behaviors; $3DWPM_1$ have also a linear comportment but is around 6 times slower (almost 2 minutes are necessary for the Feline model). The low performance of $3DWPM_1$ and $3DWPM_2$ depends mainly on their implementation that it is not optimized. If we look at the algorithmic complexity, $3DWPM_1$ and $3DWPM_2$ are linear with the number of vertices. The theoretical complexity of $MSDM$ is quadratic but it can be reduced to linear if the meshes to compare share the same connectivity and vertex order.

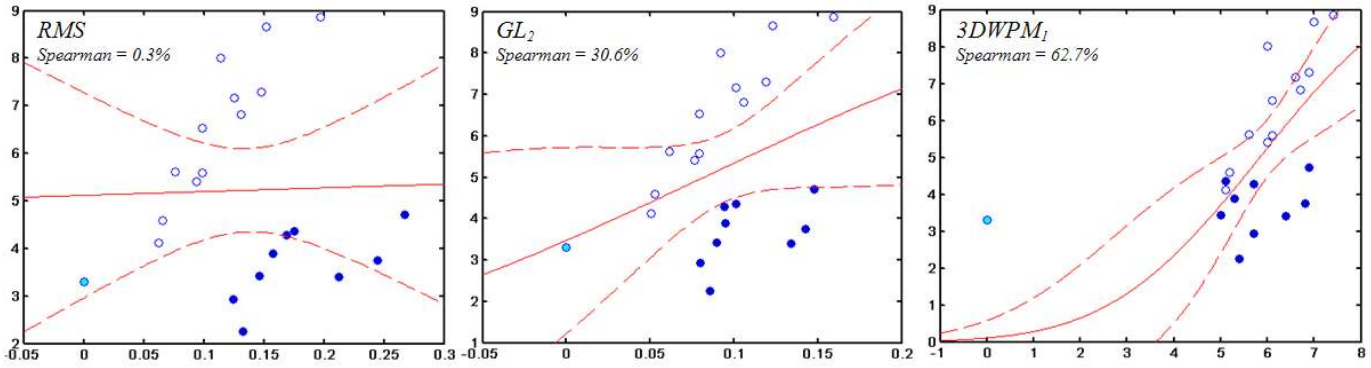

Fig. 5. Subjective MOS vs metric values for the Dyno model and for different metrics. Each circle represents a distorted model; empty circles are the models with noise and blue filled circles are the smoothed models; the original model is represented by the light blue filled circle. The Gaussian fitted curve is displayed in red, the dashed lines illustrate the confidence interval for that curve.

TABLE 4
Processing times (in seconds) of the different metrics, for 3D objects of different sizes.

	Bimba (8.8K)	Lion (38.7K)	Feline (64.5K)
$RMS/GL_1/GL_2$	< 0.2	< 0.2	< 0.2
H_d	1.6	7.2	14.2
$MSDM$	1.5	11.4	20.3
$3DWPM_1$	20.2	81.5	119.3
$3DWPM_2$	3.9	11.9	17.2
SF	< 0.2	< 0.2	< 0.2

7 Texturing and Complex Materials

This work deals with geometry rendered with standard shading algorithms, more specifically Gouraud shading with Phong lighting model, i.e. basic OpenGL rendering, considering objects made with diffuse stone-like material and illuminated by a single point light source. This is one of the main limitations of this work, texture mapping, materials with complex reflectance behavior and complex lighting environments are not taken into account in the performance evaluations of the geometry-based perceptual metrics. Obviously, the presence of these factors affect visual perception of the rendered 3D object, making the results of our metrics not completely reliable in visually rich rendering contexts. Hence, it is important to make some considerations about these aspects.

Concerning texture mapping, some previous works that deal with textured models exist, for example the masking model of Ferwerda et al. [16], the work of Qu and Meyer [11] for the perceptual remeshing of 3D textured models, the perceptual metric of Williams et al. [10] that deals also with lighting conditions, and others. We recall here, that these perceptual metrics work in image-space and not directly on the 3D model surface, making their results effective but not completely reliable. The big challenge is to account for the presence of textures by working directly on the objects' surface, for example evaluating the masking effect of the texture using both the texture coordinates and the image content of the texture map. This issue involves the adaptation of current visual masking models to the parameterized textured surface by calculating for each point of the surface, for example, a map of masking that can be used together with the geometric-based perceptual metrics to evaluate visual differences between textured 3D models. Anyway, the metrics here described, especially the MSDM one, can be seen as a lower bound concerning this visual factor. In fact, texture mapping can hide completely or partially visual changes caused by geometry changes of the surface but with very low probability it could increase the perception of such distortions. In other words, we can consider perceptually-motivated geometric metrics conservative with respect to textured objects.

Different considerations have to be drawn for materials which exhibit complex reflectance behavior, in fact, in this case the visual effect on the rendered model depends strongly on the viewpoint (considering material with strong specular reflection), hence, it is very difficult to embed this issue in a object-based metric. Some attempt to model the influences of this and other visual factors with a perceptual metric that consider the overall 3D scene are the works of Ramanarayanan et al. [17] and Ferwerda et al. [45] that presents very promising results. The open issue, from our point of view, is how to adapt such studies locally to the model surface.

Finally, 3D models are usually viewed interactively. So, the objects' movement is another important factor that the perceptual metrics should taken into account. The metric of Yee et al. [19] and Myszkowski [20] for computer animations could represent a good base for integrating also the effect of the user interaction (for example for video games applications) into model-based perceptual metrics.

For the above reasons we think that most of these aspects require a big research effort and merit separate and specific studies and that the geometry-based perceptual

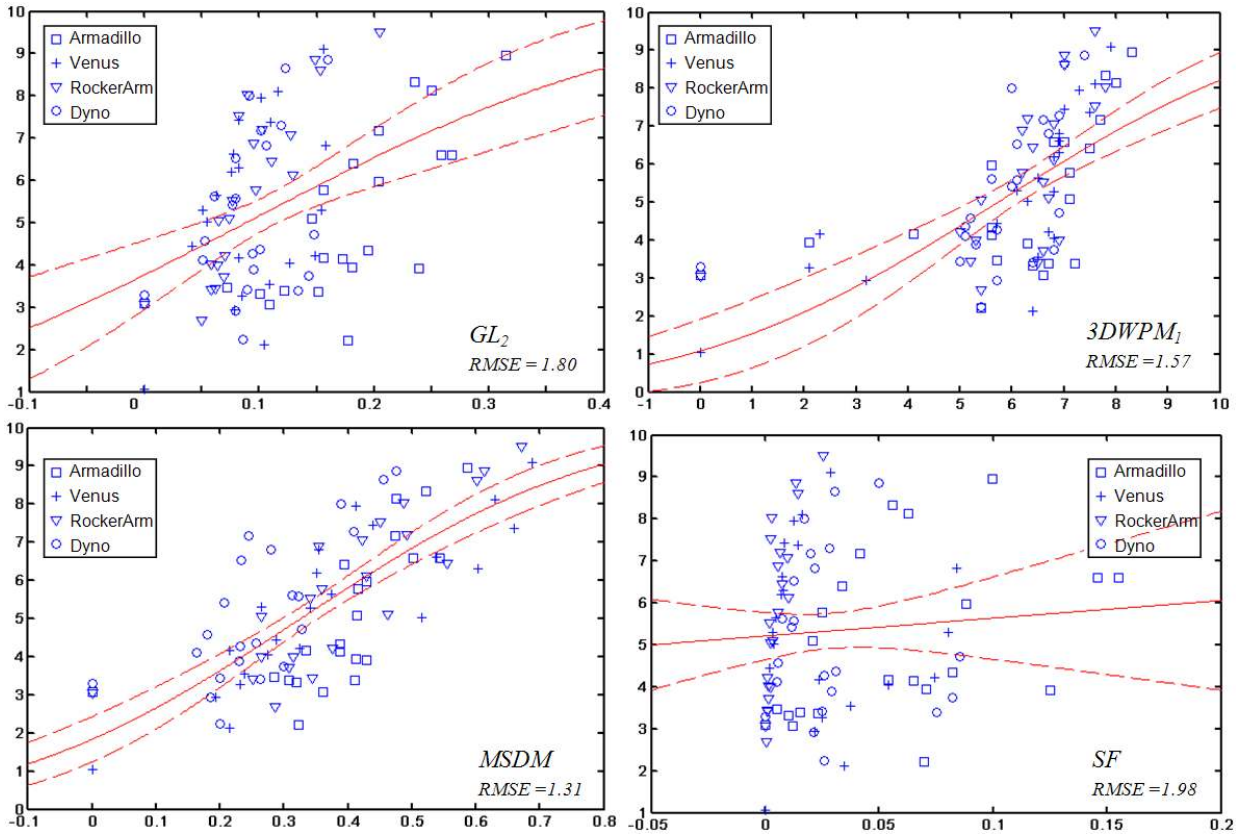

Fig. 6. Subjective MOS vs metric values for the whole corpus and for different metrics. Each symbol represents a distorted model. The Gaussian fitted curve is displayed in red, the dashed lines illustrate the confidence interval for that curve.

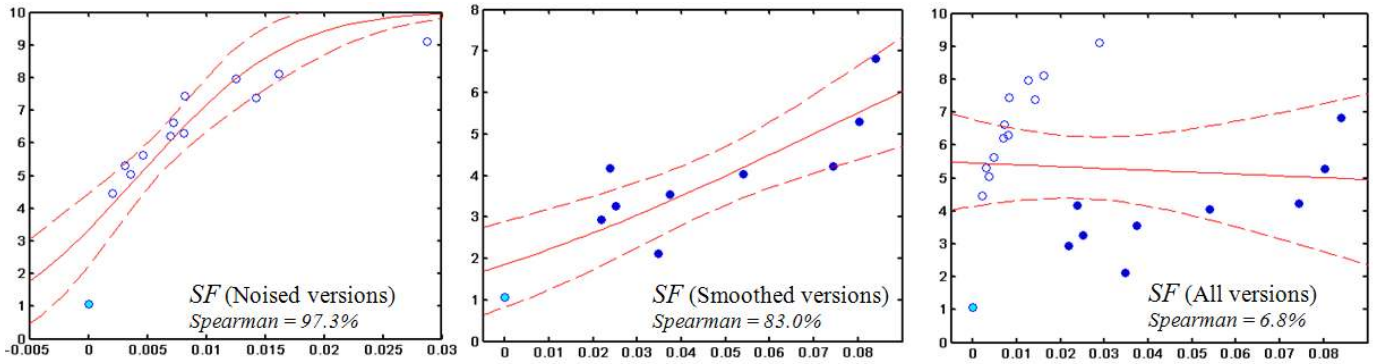

Fig. 7. Subjective MOS vs SF metric values for the RockerArm model. *Left:* only noisy versions are taken into account, *Middle:* only smoothed versions, *Right:* all distortions are taken into account.

metrics here studied are important by themselves even under the limitations just described.

8 Summary and Conclusion

We have presented quantitative and qualitative evaluations and comparisons of existing perceptually-inspired geometry-based 3D metrics for triangular meshes. Subjective experiments based on different corpuses and evaluation protocols have been considered, corresponding to different use cases and perceptual mechanisms. Several final remarks can be outlined:

- Existing model-based perceptual metrics are still not really able to take into account the masking effect and to properly handle different kinds of artifacts such as smoothing and noise addition. Despite this most of them (GL_2, $3DWPM$, $MSDM$, SF) have a good behavior regarding the frequency sensitivity of the human visual system.
- Basically, perceptually-motivated metrics ($MSDM$, $3DWPM$) perform largely better than standard geometric ones (RMS, $Hausdorff$). More precisely, $MSDM$ seems to provide the best results, with

Spearman correlations around 70% for both visual masking and general purpose corpuses. $3DWPM_1$ exhibits rather good performances on the general purpose corpus (Spearman correlation is 69%); if only noise-like artifacts are considered, $3DWPM_2$ has the best performances (87.7% for Spearman correlation on the noisy models of the general purpose corpus). Surprisingly RMS seems to well catch the visual masking effect (mean Spearman correlation per model is 70% for the Visual Masking corpus).

- No test about the comparison of mesh and its simplified version has been conducted here due to the problem to construct a representative corpus. Moreover, only one perceptual metric dealing explicitly with this kind of processing, the one of Yixin Pan et al. [25], has been developed since now. This is an important open issue since simplification algorithms are a common geometric processing operation on 3D meshes.

Concluding, in Computer Graphics, there still lacks an efficient mechanistic measure like those existing for 2D image quality assessment as the *VDP* [4] or the Sarnoff model [9]. Existing works which have attempted to generalize concepts from visual perception to computer graphics are mainly view-dependent and rely in fact on 2D image perceptual mechanisms applied on the rendered image. As we know from the study by Rogowitz and Rushmeier [18] this approach is not completely reliable. For this reason, a critical point will be the generalization of perceptual cognitive mechanisms in the 3D graphics context, in particular by considering mainly the geometry of the models. Such metrics should integrate also other aspects which influence the final visual appearance of the rendering such as lighting, texture mapping and the material properties of the model itself. This would constitute a big step in the development of really efficient perceptually-driven graphics algorithms. The work here presented aims to offer a small step in this direction.

ACKNOWLEDGMENT

We would like to thank Jérôme Lavoué for his great help on statistical analysis.
This work is partially supported by the French National Research Agency (ANR) through MADRAS project (ANR-07-MDCO-015) and by the EC IST IP project 3D-COFORM (IST-2008-231809).

REFERENCES

[1] J.-W. Cho, R. Prost, and H.-Y. Jung, "An oblivious watermarking for 3-d polygonal meshes using distribution of vertex norms," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 142–155, 2007.

[2] M. Garland and P. S. Heckbert, "Surface simplification using quadric error metrics," in *ACM Siggraph*, 1997, pp. 209–216.

[3] K. Hildebrandt and K. Polthier, "Anisotropic filtering of non-linear surface features," *Computer Graphics Forum*, vol. 23(3), pp. 391–400, September 2004, proc. Eurographics 2004.

[4] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," pp. 179–206, 1993.

[5] M. Eckert and A. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing*, vol. 70, no. 3, pp. 177–200, 1998.

[6] Z. Wang and A. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[7] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to jpeg2000," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, 2004.

[8] P. Lindstrom and G. Turk, "Image-driven simplification," *ACM Trans. Graph.*, vol. 19, no. 3, pp. 204–241, 2000.

[9] J. Lubin., "A visual discrimination model for imaging system design and evaluation," pp. 245–283, 1995.

[10] N. Williams, D. Luebke, J. D. Cohen, M. Kelley, and B. Schubert, "Perceptually guided simplification of lit, textured meshes," in *I3D '03: Proceedings of the 2003 symposium on Interactive 3D graphics*. New York, NY, USA: ACM, 2003, pp. 113–121.

[11] L. Qu and G. W. Meyer, "Perceptually guided polygon reduction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 5, pp. 1015–1029, 2008.

[12] M. Reddy, "Perceptually modulated level of detail for virtual environments," Ph.D. dissertation, Dept. of Computer Science, University of Edinburgh, UK, 1997.

[13] M. R. Bolin and G. W. Meyer, "A perceptually based adaptive sampling algorithm," in *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM, 1998, pp. 299–309.

[14] M. Ramasubramanian, S. N. Pattanaik, and D. P. Greenberg, "A perceptually based physical error metric for realistic image synthesis," in *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 73–82.

[15] R. Dumont, F. Pellacini, and J. A. Ferwerda, "Perceptually-driven decision theory for interactive realistic rendering," *ACM Trans. Graph.*, vol. 22, no. 2, pp. 152–181, 2003.

[16] J. A. Ferwerda, P. Shirley, S. N. Pattanaik, and D. P. Greenberg, "A model of visual masking for computer graphics," in *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 143–152.

[17] G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala, "Visual equivalence: towards a new standard for image fidelity," in *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*. New York, NY, USA: ACM, 2007, p. 76.

[18] B. E. Rogowitz and H. E. Rushmeier, "Are image quality metrics adequate to evaluate the quality of geometric objects," in *Human Vision and Electronic Imaging*, 2001, pp. 340–348.

[19] H. Yee, S. Pattanaik, and D. P. Greenberg, "Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments," *ACM Trans. Graph.*, vol. 20, no. 1, pp. 39–65, 2001.

[20] K. Myszkowski, "Perception-based global illumination, rendering, and animation techniques," in *SCCG '02: Proceedings of the 18th spring conference on Computer graphics*. New York, NY, USA: ACM, 2002, pp. 13–24.

[21] S. Kim, S. Kim, and C. Kim, "Discrete differential error metric for surface simplification," in *Pacific Graphics*, 2002, pp. 276–283.

[22] S. Howlett, J. Hamill, and C. O'Sullivan, "An experimental approach to predicting saliency for simplified polygonal models," in *APGV '04: Proceedings of the 1st Symposium on Applied perception in graphics and visualization*. New York, NY, USA: ACM, 2004, pp. 57–64.

[23] C. H. Lee, A. Varshney, and D. W. Jacobs, "Mesh saliency," in *SIGGRAPH '05: ACM SIGGRAPH 2005 Papers*. New York, NY, USA: ACM, 2005, pp. 659–666.

[24] Z. Karni and C. Gotsman, "Spectral compression of mesh geometry," in *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 279–286.

[25] Y. Pan, I. Cheng, and A. Basu, "Quality metric for approximating subjective evaluation of 3-d objects," *Multimedia, IEEE Transactions on*, vol. 7, no. 2, pp. 269–279, April 2005.

[26] E. Drelie Gelasca, T. Ebrahimi, M. Corsini, and M. Barni, "Objective Evaluation of the Perceptual Quality of 3D Watermarking," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2005.

[27] M. Corsini, E. Drelie Gelasca, T. Ebrahimi, and M. Barni, "Watermarked 3d mesh quality assessment," *IEEE Transaction on Multimedia*, vol. 9, no. 2, pp. 247–256, February 2007.
[28] G. Lavoué, E. Drelie Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, "Perceptually driven 3d distance metrics with application to watermarking," in *SPIE Applications of Digital Image Processing XXIX*, vol. 6312, 2006.
[29] Z. Bian, S.-M. Hu, and R. Martin, "Comparing small visual differences between conforming meshes," in *GMP*, ser. Lecture Notes in Computer Science, F. Chen and B. Jüttler, Eds., vol. 4975. Springer, 2008, pp. 62–78.
[30] Z. Bian, S.-M. Hu, and R. R. Martin, "Evaluation for small visual difference between conforming meshes on strain field," *Journal of Computer Science and Technology*, vol. 24, no. 1, pp. 65–75, 2009.
[31] O. Sorkine, D. Cohen-Or, and S. Toldeo, "High-pass quantization for mesh encoding," in *Eurographics Symposium on Geometry Processing*, 2003, pp. 42–51.
[32] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro : Measuring error on simplifed surfaces," *Computer Graphics Forum*, vol. 17, no. 2, pp. 167–174, 1998.
[33] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from errorvisibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 1–14, 2004.
[34] J.-H. Wu, S.-M. Hu, J.-G. Sun, and C.-L. Tai, "An effective feature-preserving mesh simplification scheme based on face constriction," *Computer Graphics and Applications, Pacific Conference on*, vol. 0, p. 0012, 2001.
[35] M. Corsini, E. Drelie Gelasca, and T. Ebrahimi, "A Multi-Scale Roughness Metric for 3D Watermarking Quality Assessment," in *Workshop on Image Analysis for Multimedia Interactive Services 2005, April 13-15, Montreux, Switzerland.*, ser. ISCAS. SPIE, 2005.
[36] R. Ohbuchi, A. Mukaiyama, and S. Takahashi, "A frequency-domain approach to watermarking 3d shapes," *Computer graphic forum*, vol. 21, no. 3, pp. 373–382, 2002.
[37] K. Wang, G. Lavoué, F. Denis, and A. Baskurt, "Hierarchical watermarking of semi-regular meshes based on wavelet transform," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 620–634, 2008.
[38] B. Vallet and B. Lévy, "Spectral geometry processing with manifold harmonics," *Computer graphic forum*, vol. 27, no. 2, pp. 251–260, 2008.
[39] G. Lavoué, "A local roughness measure for 3d meshes and its application to visual masking," *ACM Transactions on Applied Perception*, vol. 5, no. 4, p. 21, 2009.
[40] E. D. Gelasca, "full-reference objective quality metrics for video watermarking, video segmentation and 3d model watermarking," Ph.D. dissertation, EPFL, 2005.
[41] "ITU Recommendation BT.500-10: Methodology for subjective assessment of the quality of television pictures," 2000.
[42] J. L. Shrout, P. and Fleiss, "Intraclass correlation: uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.
[43] W. W. Daniel, *Biostatistics: A Foundation For Analysis In The Health Sciences Books, 7th edition*. John Wiley and sons, 1999.
[44] G. Taubin, "A signal processing approach to fair surface design," in *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM, 1995, pp. 351–358.
[45] J. Ferwerda, G. Ramanarayanan, K. Bala, and B. Walter, "Visual equivalence: an object-based approach to image quality." in *Proceedings IS&T 16th Color Imaging Conference*, 2008.

Guillaume Lavoué received the engineering degree in signal processing and computer science from CPE Lyon (2002), the MSc degree in image processing from the University Jean Monnet, St-Etienne (2002) and the PhD degree in computer science from the University Claude Bernard, Lyon, France (2005). After a postdoctoral fellowship at the Signal Processing Institute (EPFL) in Switzerland, since September 2006 he has been an associate professor at the French engineering university INSA of Lyon, in the LIRIS Laboratory (UMR 5205 CNRS). His research interests include 3D model analysis and processing, including compression, watermarking, perception, and 2D/3D recognition.

Massimiliano Corsini received the Degree (Laurea) in Information Engineering from University of Florence. In 2005 he received a PhD degree in Information and Telecommunication Engineering from the same University working on 3D watermarking of polygonal meshes and perceptual metrics for 3D watermarking quality assessment. Currently, he is a Researcher at the Institute of Information Science and Technologies (ISTI) of the National Research Council (CNR) in Pisa, Italy. His research interests are in the fields of Computer Graphics, Computer Vision and Image Processing and include 3D watermarking, perceptual metrics, visual appearance acquisition and modeling and image-based relighting.