# Perceptual Quality Assessment of 3D Dynamic Meshes:<br>Subjective and Objective Studies

Fakhri Torkhani, Kai Wang*, Jean-Marc Chassery

*GIPSA-Lab, UMR 5216 CNRS, University of Grenoble Alpes,*
*11, rue des Mathématiques, F-38402 Saint Martin d'Hères Cedex, France*

**Abstract**

Nowadays, 3D mesh animations have been increasingly used in various applications, *e.g.*, in digital entertainment and physically-based simulation. In many applications, it is possible that a surface animation undergoes some lossy operations which can impair its perceptual quality. Since the end users of mesh animations are often human beings, the perceptual quality assessment of 3D dynamic meshes has recently emerged as an interesting and important research problem. In this paper, we will first of all present the results of a large-scale subjective experimental study carried out to collect human opinion scores of the perceptual quality of distorted dynamic meshes. We integrate in our subjective study a large set of 276 impaired mesh sequences with different distortion intensities and categories, such as spatial and temporal visual masking simulation, lossy compression and network transmission error. The constructed database is publicly available and can serve as ground-truth data for the performance evaluation of objective metrics of dynamic mesh perceptual quality. A comparative study is then conducted to benchmark existing objective metrics. The main finding of this objective study is that in general model-based quality metrics outperform image- and video-based quality metrics, but none of the existing metrics can achieve high correlation with human opinion scores on the whole subjective database and for all kinds of distortions. Finally, a simple but effective full-reference objective metric is proposed for the perceptual quality assessment of 3D dynamic meshes, which incorporates both spatial and temporal features and which provides a satisfactory correlation value when tested on the subjective database.

*Keywords:* Dynamic mesh, 3D animation, perceptual quality assessment, subjective experiment, objective metric

## 1. Introduction

In many graphics and multimedia applications, 3D objects are digitally represented by triangle meshes [1]. A static triangle mesh can be fully described by two kinds of information: the 3D coordinates of mesh vertices (*i.e.*, the *geometry* information) and the adjacency relationship between vertices (*i.e.*, the *connectivity* information). A dynamic mesh is composed of a sequence of static meshes, and is commonly used to represent an object whose shape and/or position changes over time. For example, a mesh sequence can represent a virtual character in a video game, a piece of tissue in physically-based simulation, or a human body in movement reconstructed from data captured from real world. It is quite possible that dynamic meshes are subject to lossy operations such as compression and watermarking, and are transmitted over error-prone channels. Such operations and transmission can introduce distortion to the geometry (*i.e.*, 3D shape) and the motion of the surface animation and therefore impair its perceptual quality. In this study, we will focus on the perceptual quality assessment (PQA) of *constant-connectivity* mesh sequences, *i.e.*, sequences whose number of vertices is invariant over time and whose vertices are always connected to the same neighbors.

A possible way to measure the perceptual quality of dynamic meshes is to average *subjective* quality scores provided by a group of human observers. Human judgments can serve as ground-truth data of dynamic mesh perceptual

*Corresponding author

*Email addresses:* fakhri.torkhani@gipsa-lab.grenoble-inp.fr (Fakhri Torkhani), kai.wang@gipsa-lab.grenoble-inp.fr (Kai Wang), jean-marc.chassery@gipsa-lab.grenoble-inp.fr (Jean-Marc Chassery) quality; however, in most cases subjective assessment cannot be a practical solution because collecting subjective scores is time consuming and costly. Alternatively, *objective* metrics provide the possibility to automate the perceptual quality assessment. Hence, they can be easily deployed along with various dynamic mesh processing algorithms, to assess and further control the resulting perceptual quality after lossy processing. The main challenge for objective metrics is to provide quality measures as close as possible to mean subjective scores [2, 3]. Ground-truth subjective scores are therefore essential to evaluate the performance of any new objective metric. To address this need, in the first part of this paper we focus on the construction of a new subjectively-rated database of 3D dynamic meshes. In order to construct a general-purpose database on which objective metrics can be reliably evaluated and compared, we include a large set of impaired mesh sequences obtained after applying various lossy operations. The constructed database comprises 276 impaired surface animations derived from 10 intact, *reference* dynamic meshes, using different distortion types of different intensities.

In our subjective study, each impaired or reference dynamic mesh was inspected by 25 observers in a controlled laboratory environment. The observers can freely interact with the mesh sequence (*i.e.*, zooming, rotation and translation of the mesh sequence using a mouse). In order to study the effect of user interaction on the subjective quality assessment result, and more importantly to easily and accurately evaluate image- and video-based PQA metrics when applied on 3D dynamic meshes (explained in the next paragraph), each mesh sequence in the database was further subjectively rated by 16 additional observers under a predefined fixed viewpoint, without the possibility to interact with the animation. This predefined viewpoint was carefully selected as the one from which we can observe as much information of the mesh sequence as possible, *i.e.*, a kind of "best view" of the dynamic mesh. We observe that under most distortion types, there is significant difference between the mean subjective scores collected from these two different experimental settings.

Based on the collected subjective scores, in the second part of this paper we perform a comprehensive and statistical comparison of the performance of the latest objective quality metrics, including metrics for 3D static and dynamic meshes (*model-based* metrics), and metrics for 2D images and videos applied for 3D dynamic mesh PQA (*image-* and *video-based* metrics). For image- and video-based metrics, we apply them on the video of surface animation recorded under the aforementioned fixed viewpoint. For metrics designed for static meshes and 2D images, we apply them on a frame-by-frame basis and take the average of scores on individual frames as the final metric output. In general, it is demonstrated that current image- and video-based metrics are not suitable to evaluate the perceptual quality of 3D dynamic meshes, even under a fixed viewpoint. Model-based metrics, designed specifically for PQA of either static or dynamic meshes, outperform image- and video-based metrics, but there is no objective metric that can provide satisfying performance on the whole subjective database and for all types of distortions.

In light of the performance evaluation results of the existing metrics, we propose in the last part of this paper a new objective metric for 3D dynamic mesh PQA. The proposed metric makes use of both spatial and temporal features. The spatial feature describes the local roughness of the 3D surface, while the temporal features are related to the speed and the motion direction of mesh vertices. The performance of the proposed metric is evaluated on the constructed subjective database. Experimental results show that in order to achieve a high correlation value with subjective scores (on the whole database, as well as for each kind of distortion), it is necessary to incorporate in an objective metric some perceptually relevant temporal features.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the state of the art in subjective and objective PQA of 3D meshes. In Section 3 we present the details of our large-scale subjective experiments aiming at collecting human opinion scores on the perceptual quality of impaired dynamic meshes. Section 4 provides a detailed analysis of the results of our subjective experiments. Based on the ground-truth subjective values collected in Section 3, we carry out in Section 5 a comprehensive comparative study of the performance of existing model-based, and image- and video-based objective metrics. To our knowledge, this comparative objective study is the first attempt of this type in the context of 3D dynamic mesh PQA. In Section 6, a new dynamic mesh PQA metric is proposed, and its performance evaluation results on the constructed subjective database are also provided. We draw conclusions and suggest several future working directions in Section 7.

## 2. Perceptual Quality Assessment of 3D Meshes

### 2.1. Prior art

Recently, image and video PQA research community has made considerable efforts on the design and construction of subjectively-rated perceptual quality databases. Several documentations, in particular the recommendations from ITU [4, 5], have been established and used to define experimental settings and protocols of such subjective experiments on images and videos. Many subjectively-rated databases have been released during last years. Winkler [6] provided an overview and a comparison of the publicly available image and video PQA databases.

First studies on subjective quality assessment of 3D static meshes were led by Watson *et al.* who tried to measure the visual fidelity of simplified meshes [7], followed by Corsini *et al.* who attempted to accustom existing experimental protocols to the subjective quality assessment of watermarked 3D meshes [8]. Recently, several publicly available databases of static meshes with associated *mean opinion scores* (MOS) have been released. The *LIRIS/EPFL General-Purpose Database* [9] contains 84 impaired models derived from 4 reference meshes. The included distortions are noise addition and smoothing applied in different regions of the mesh surface. The *LIRIS Masking Database* [10] was created to study the spatial visual masking effect. This database includes 24 impaired models derived from 4 reference meshes. Models were carefully selected to offer broad range of roughness, and noise was added in either rough or smooth regions. The *IEETA Simplification Database* [11] includes 30 simplified models (obtained by using different vertex reduction algorithms) derived from 5 reference meshes. The *UWB Compression Database* [12] contains 63 impaired, geometrically-compressed meshes derived from 5 reference models. The above databases have been used to evaluate and compare the most recent objective perceptual quality metrics for 3D static meshes [3], such as $MSDM2$ [13], $DAME$ [12], $FMPD$ [14] and $TPDM$ [15].

To our knowledge, the *UWB Dynamic Mesh Database* [16] is the first and only existing subjective quality database of 3D dynamic meshes. This database includes 36 impaired mesh sequences derived from 4 reference surface animations. Considered distortions include various kinds of Gaussian noises, random uniform noise, smooth sinusoidal noise and geometry compression distortions. Classical geometric distance metrics (*e.g.*, the $KG$ metric [17]) and the perceptually-driven metric $STED$ [16] have been tested and compared on this database.

### 2.2. Motivations and objectives

As pointed out in [18], for reliable evaluation and comparison of objective PQA metrics, we need a subjectively-rated database composed of a large number of distorted visual contents. This conclusion is backed by rigorous proof based on the level of significance of statistical hypothesis testing [18]. In practice, the number of distorted contents is more than 150 in the case of video PQA study [19, 20, 21, 22], a close but much more mature research area than PQA of 3D dynamic meshes. On the contrary, the only existing dynamic mesh perceptual quality database [16] comprises only 36 distorted models. One more limitation of this database is that there is no consistency between MOS values of distorted models derived from two different reference dynamic meshes. Therefore, the database in [16] can, to some extent, be considered as 4 small datasets of 9 distorted models, because the correlation between objective and subjective scores on the whole database is not a reliable performance benchmarking value. This is actually in accord with the experimental results shown in the original paper of this database [16], where the authors reported 4 correlation values for each group of 9 distorted models derived from a same reference dynamic mesh, without providing the overall correlation on the whole database. Besides the low number of distorted models per reference mesh, the database also has a low number of impaired models per type of distortion. For each kind of distortion, at most only 1 impaired model was generated per reference mesh. Therefore, in all, for each kind of distortion, there are at most 4 impaired models included in the database. This makes very difficult to reliably evaluate the performance of objective metrics in assessing the perceptual quality of dynamic meshes impaired by a given type of distortion.

Hence, considering the above observations, our first objective in this paper is to construct a new subjectively-rated database comprising a large set of impaired dynamic meshes. The construction of such a large-scale subjective database is expected to fulfill the urgent requirement of reliable ground-truth data for fair evaluation and comparison of objective metrics. The subjective experiments should follow as closely as possible the relevant, well-established ITU recommendations [4, 5]. Meanwhile, we should also ensure the consistency between subjective scores of models impaired by different types of distortions and derived from different reference dynamic meshes. Through the subjective experiments, we also want to conduct a first trial in the literature on the study of *spatial and temporal visual masking effects* in the case of 3D dynamic mesh PQA. It is believed that a good objective metric should be able to capture these important visual masking effects. Finally, it is worth mentioning that the constructed database, the associated subjective scores and the experiment software are freely shared on-line[1]. With the constructed subjective database, our second and third objectives naturally consist in the performance evaluation and comparison of existing objective metrics and the development of a more effective model-based metric.

## 3. Subjective Study

### *3.1. Assessment method*

Since there is no specific recommendation on designing subjective quality assessment experiments on 3D dynamic meshes, we have made an effort to accustom existing ITU recommendations on images and videos [4, 5, 23] to establish an appropriate assessment method for 3D dynamic models. Standardized assessment methods are characterized by how stimuli are presented. Within *Double Stimulus* (DS) method, two stimuli are simultaneously presented to observers, and the viewer's task is to judge and quantify the quality of the impaired stimulus given the reference signal. In our subjective study, we opted for a *Single Stimulus* (SS) method: each time only a single dynamic mesh is shown to the observer and the viewer's task is to give a score reflecting the perceptual quality of the observed stimulus. Test materials include impaired mesh sequences with randomly inserted intact hidden reference sequences. Hidden reference sequences are presented as any other test stimulus. Apart from the considerable saving of testing time compared with DS method [18, 19], another advantage of SS method is that it allows future use of the database to investigate issues related to the measurement of intrinsic quality of dynamic meshes and to facilitate the evaluation of *no-reference*[2] objective metrics. Moreover, in SS method the quality degradation of an impaired mesh can be computed as the difference of the two scores assigned respectively to the hidden reference model and to the impaired model. This differential quality score allows us to easily evaluate the performance of *full-reference* objective metrics.

Perceptual quality assessment of 3D dynamic meshes is a complex problem with many influencing factors, *e.g.*, modifications in vertex coordinates, lighting, surface texture and color, surface material, camera position, *etc.* Among these factors, vertex coordinate modification is particularly important, and it can introduce distortions in both shape and motion perception. In our study, we will concentrate on this factor. Our strategy is to generate various impaired meshes by modifying vertex positions and show the impaired meshes to observers under common rendering and viewing conditions. Extension of the database to study other influencing factors is possible in the future. At present, test conditions used in our subjective experiments are as follows.

**Shading.** Animation surfaces are naturally curved. To avoid the visual effect of triangular sampling on 3D surfaces, we decided to render dynamic meshes using the well-known and commonly used *Gouraud* smooth shading technique.

**Surface material and color.** Surfaces were considered as diffusive materials without any specular reflection to reduce the viewpoint dependency [8]. Similar to the study in [8], a non-uniform blue-to-white color background was used, so as to avoid exaggerating the importance of model silhouette on the result of quality assessment. As mentioned above, we focus on the visual distortion caused by the modification of vertex coordinates, therefore we did not use any mapped texture on mesh surfaces and set the surface color as mid gray.

**Lighting.** A static light source was used. Lighting orientation was set along the normal of the screen to fit the observer's viewing direction.

**User interaction.** Two categories of experiments were conducted, *i.e.*, *with-user-interaction* and *without-user-interaction* experiments. During with-user-interaction experiments observers could freely zoom, rotate and translate the dynamic mesh under evaluation. Observer interactions during the experiments were all saved for a future study. For without-user-interaction experiments, test sequences were displayed under a fixed viewpoint, so they could be considered like video tracks. As mentioned earlier, we decided to explore these two different experimental settings for two reasons: 1) to investigate the influence of user interaction on the perceived quality and the resulted subjective scores; 2) to easily and accurately evaluate the performance of image- and video-based objective metrics.

[1] Data and software are available at `http://www.gipsa-lab.fr/~kai.wang/software/database/`. The password to unzip the archives is GIPSA3DMAQD.

[2] Objective PQA metrics can be classified according to the amount of information available about the reference source that is assumed to be of perfect quality, as full-reference metrics (*i.e.*, the reference content is completely available), no-reference metrics (*i.e.*, no information is available about the reference content), or reduced-reference metrics (*i.e.*, part of the information about the reference content is available).

**Camera.** The same initial viewpoint (*i.e.*, the initial camera position and direction when starting an animation) was preselected and assigned to all observers. The initial viewpoint was carefully selected with the objective to best present the spatio-temporal and semantic content of the animation. For the same purpose, we used in the experiments two modes of cameras: static camera and embedded (object-tracking) camera. For mesh sequences with movement towards the depth of the scene an embedded camera was used which was toggled to the surface animation, and the static camera was used for other sequences. As explained above, in with-user-interaction experiments observers can freely change the viewpoint.

**Environment setting.** Experiments were organized in a laboratory controlled environment. Meshes were displayed on a TFT-LCD screen of 22 inch size, in a low illumination room. Screen resolution was $1680 \times 1050$, a viewing distance of 40 *cm* was preserved for all participants in our experiments.

**Rating.** For the perceptual quality rating, we adopted an absolute numerical categorical rating scale with 11 grades (from 0 to 10). The rating scale was sampled to five overall quality levels labeled as: bad, poor, fair, good and excellent. A prior ITU study described in [24] showed the appropriateness of this method when applied in SS experiments.

**Duration.** Each mesh sequence was shown to observers for 30 seconds in the with-user-interaction tests and 25 seconds in the without-user-interaction tests. During this period, the animation was repeatedly played in a forward-backward cyclic manner. We organized the experiments in six sessions: three with user interaction and the other three without user interaction. Following the ITU recommendation [4] and to avoid observer fatigue, sessions with viewing duration exceeding 30 minutes were divided into two sub-sessions. Observers took a break of 2 minutes between sub-sessions.

**Training.** The training session attempted to familiarize observers with dynamic meshes whilst avoiding any memory effect. Four other mesh sequences not included in the real tests were selected for being used in the training session. The training test materials were the same for all experimental sessions. Impaired training materials were obtained after applying random uniform and Gaussian additive noises. Observers were informed that they would not necessarily see the same type of distortion during the real tests. To help participants understand their task, a message was shown when displaying animations with the highest noise intensity to inform observers that they were watching an example animation with very bad quality. Similarly, a message was shown to inform participants that they were watching an animation with very high quality when displaying intact reference sequences.

**Software.** An OpenGL-based software tool was developed for the subjective experiments. To avoid buffering problem, the next mesh sequence to be evaluated was loaded to the computer memory while displaying the current sequence. This was realized via multi-thread programming. Necessary processing for the animation display such as normal direction computation was accomplished before starting the animation. The subjective assessment interface is shown in Fig. 1. It comprises a stimulus presentation window along with a horizontal bar marked with graduated numerical-categorical rating scales. Observers had the task of observing the dynamic mesh (with or without interaction), establishing an opinion about the overall quality and pressing vote button to proceed to the next mesh sequence. Observers were free to vote before the end of the viewing duration to proceed to the next sequence, but they did not have the possibility to view the same animation twice. For with-user-interaction sessions, observers could at any time reset the camera to the preselected initial viewpoint, by clicking a button specifically designed for this purpose.

### *3.2. Observers*

For the three with-user-interaction sessions, each mesh sequence was evaluated by 25 observers. For the three without-user-interaction sessions, each sequence was evaluated by 16 observers. No observer evaluated the same impaired mesh animation under both with- and without-user-interaction sessions. Observers, aged between 18 and 58, were students and staffs from the University of Grenoble Alpes, and they were inexperienced in mesh perceptual quality assessment. Before experiments, observers were screened by a test of normal or corrected-to-normal vision acuity using a *Snellen* chart.

The experiment involved five stages: 1) reading written instructions, 2) oral instructions, 3) training, 4) test session, and 5) interview. The first two steps aimed to inform observers about the purpose of the experiment and the instructions to follow. The short training stage was necessary to help observers get familiar with the test conditions, the distortion range and the assessment software. In addition, 5 “dummy” animations were included at the beginning of each test session, with the objective to stabilize the opinion of observer. Scores collected from “dummy” animations will not be

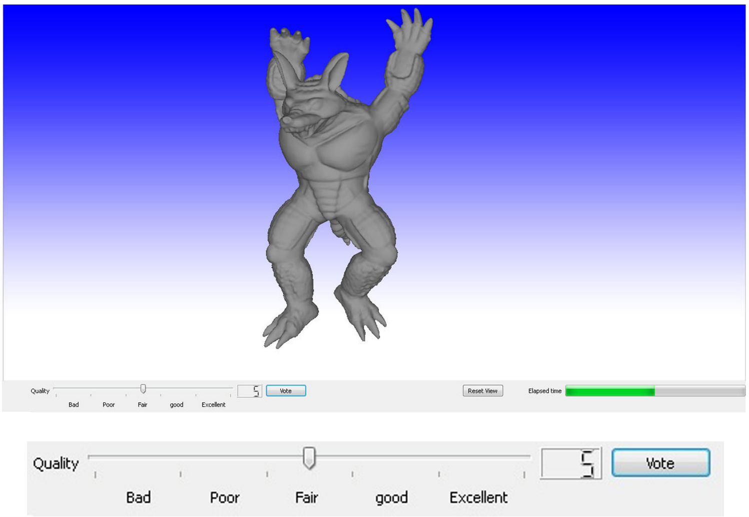


Figure 1: The software interface of subjective experiments, with a close-up image (below) of the horizontal rating scale bar.

taken into account in the result processing and analysis. In order to avoid any memory-based bias, mesh animations in each session for each observer were displayed in a pseudo-random order. Furthermore, two mesh sequences derived from the same reference animation were not allowed to be displayed one after another.

### *3.3. Test materials*

Selecting appropriate reference contents is very important in the construction of a new subjectively-rated perceptual quality database. As mentioned in Section 1, in this study we focus on the perceptual quality assessment of constant-connectivity mesh sequences, the most commonly used dynamic models in practical applications. The number of vertices and their adjacency relationship (*i.e.*, the connectivity information) are kept invariant for all the frames in a mesh sequence. As shown in Table 1, 10 reference dynamic meshes are included in our subjective database:

- Horse and Elephant[3] are dynamic meshes representing a running animal, with smooth surface and rich temporal content. These two animations were created using the mesh deformation transfer technique proposed in [25].
- Chicken[4], from the animated film "Chicken Crossing", is a dynamic mesh with both slow and very fast movements.
- Chinchilla[5], a virtual character from the Blender computer graphics movie "Big Buck Bunny", is a mesh animation with fast local and global motions.
- Dress[6] sequence represents a smooth one-piece dress with slow movement. The geometry and motion of this sequence was captured from a real-world garment [26].
- ClothBall and Balls[7] are physically simulated dynamic meshes used in collision detection research [27, 28].
- From the static meshes Human[8], Dinosaur[9] and Armadillo[10], we generated three mesh animations using the automatic rigging and animation tool of [29].

[3] http://people.csail.mit.edu/sumner/research/deftransfer/
[4] http://www.glassner.com/creative/films-and-games/
[5] http://www.bigbuckbunny.org
[6] http://www.cs.ubc.ca/labs/imager/tr/2008/MarkerlessGarmentCapture/data.html
[7] http://gamma.cs.unc.edu/DYNAMICB/
[8] http://www.turbosquid.com/3D-models/human-base-obj-free/483277
[9] http://cyberware.com/products/scanners/desktopSamples.html
[10] http://graphics.stanford.edu/data/3Dscanrep/

Table 1: Reference dynamic meshes included in the subjective experiments.

| Sequence | # Frames | # Vertices | Camera |
|---|---|---|---|
| Horse | 47 | 8431 | S |
| Chicken | 321 | 3030 | E |
| Chinchilla | 83 | 4307 | S |
| Elephant | 48 | 42321 | S |
| Dress | 81 | 41057 | S |
| Human | 161 | 18890 | E |
| ClothBall | 68 | 46598 | S |
| Balls | 40 | 73960 | S |
| Dinosaur | 151 | 20218 | E |
| Armadillo | 74 | 40002 | E |

S: Static camera. E: Embedded (object-tracking) camera.

We can see that the included sequences have been created using different techniques and utilized in different application fields. This is a desired property for a general-purpose PQA database. Besides, it is expected that the included sequences have different spatio-temporal characteristics [20], *i.e.*, they should cover a large range of both *Spatial Information* ($SI$) and *Temporal Information* ($TI$). Although several definitions exist for $SI$ and $TI$ of images and videos [5, 30], the computation of $SI$ and $TI$ for 3D meshes still remains an open research problem. As by-products of the subjective database, here we propose two simple definitions for $SI$ and $TI$ of 3D mesh animations. For $SI$, first of all we find it difficult to derive a proper definition that works consistently well for both coarse and dense meshes. The same effect occurs for $SI$ of images whose value generally increases as image resolution decreases [30]. Therefore, in order to reduce mesh density dependency, we decided to evaluate $SI$ on refined meshes of approximately the same number of vertices, obtained after one or more iterations of mid-point subdivision of the original mesh. This simple subdivision scheme perfectly preserves the mesh shape, and in practice we set a simple stop criterion as that the final refined mesh should have more than 50000 vertices. $SI$ is then computed on the refined mesh as:

$$SI = \frac{1}{n_f}\sum_{i=1}^{n_f}\left(\frac{1}{n_v^{'}}\sum_{j=1}^{n_v^{'}} LR_{ij}\right), \tag{1}$$

where $n_f$ is the number of frames in the animation, $n_v^{'}$ is the number of vertices in each refined frame, and $LR_{ij}$ denotes the local roughness value on vertex $v_{ij}$ (*i.e.*, $j$-th vertex in $i$-th frame) which is proposed in [14]. The local roughness measure $LR_{ij}$ is defined as the Laplacian of Gaussian curvature on each vertex and corresponds well to the human perception (see [14] for details). Similarly, the temporal information $TI$ is defined as:

$$TI = \frac{1}{n_f}\sum_{i=1}^{n_f}\left(\frac{1}{n_v^{'}}\sum_{j=1}^{n_v^{'}} S_{ij}\right), \tag{2}$$

where $S_{ij}$ is the mean amplitude of the backward and forward motion vectors of vertex $v_{ij}$, *i.e.*, $TI$ is computed in a temporal window of 3 frames. For vertices in the first (respectively last) frame, we only consider the forward (respectively backward) motion vector.

$SI/TI$ pairs of the reference dynamic meshes are plotted in Fig. 2, which shows that the sequences in our database have a good dispersity on the $SI/TI$ plane and that they cover a large range of both spatial and temporal information. $SI/TI$ values are rather consistent with human judgments (please see the recorded videos of dynamic meshes on the database website). For example, the Dress sequence represents a smooth garment with very slow movement, so it is at the left-bottom corner of the $SI/TI$ plane. On the contrary, the Armadillo sequence represents a jumping animal with bumpy and rough surface, so it is located towards the top-right conner of the plot. Snapshots of the reference animations are shown in Fig. 3. As a first trial to derive $SI$ and $TI$ metrics for 3D meshes, here we have adopted the simple non-adaptive mid-point subdivision, which multiplies the number of facets by 4 in each iteration, to mitigate the problem of dependency on mesh density. In the future, more effective $SI$ and $TI$ metrics can be developed, for example, by using adaptive subdivision and edge splitting schemes, or by deriving relevant features that are not sensitive to the mesh density.

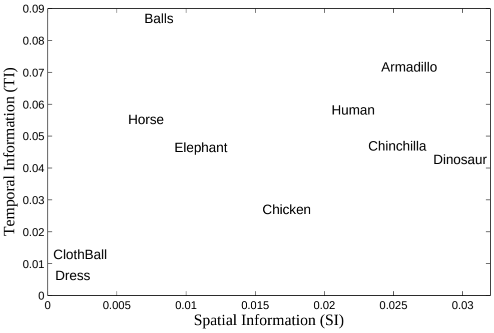


Figure 2: Spatial Information ($SI$) and Temporal Information ($TI$) of the reference sequences included in the subjective database.

Table 2: Distortions applied on the reference dynamic meshes. $SM_1$ and $SM_2$ are spatial visual masking distortions. $TM_1$ and $TM_2$ are temporal visual masking distortions. $WR$ and $IWR$ refer respectively to noise addition whose amplitude is weighted or inversely weighted by the local surface roughness. $WS$ and $IWS$ refer respectively to noise addition whose amplitude is weighted or inversely weighted by the speed of vertex. '3' means that distortions of 3 different intensity levels were generated, while '–' means that the distortion type in the corresponding column was not applied on the animation in the corresponding row.

| Animation | Simulated distortions | | | | | | | | | | Real-world distortions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Global distortions | | Descriptor-weighted distortions | | | | | | | | Compression | | | Network error |
| | Uniform | Gaussian | $SM_1$ | | $SM_2$ | | $TM_1$ | | $TM_2$ | | FAMC | | CODDYAC | |
| | | | $WR$ | $IWR$ | $WR$ | $IWR$ | $WS$ | $IWS$ | $WS$ | $IWS$ | $DCT$ | $Lifting$ | | |
| Horse | 3 | 3 | – | – | – | – | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Chicken | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | – | – | 3 | 3 | 3 | 3 |
| Chinchilla | 3 | 3 | – | – | – | – | – | – | – | – | 3 | 3 | 3 | 3 |
| Elephant | 3 | 3 | – | – | – | – | – | – | 3 | 3 | 3 | 3 | 3 | 3 |
| Dress | 3 | 3 | – | – | – | – | – | – | – | – | 3 | 3 | 3 | 3 |
| Human | 3 | 3 | 3 | 3 | 3 | 3 | – | – | – | – | 3 | 3 | 3 | 3 |
| ClothBall | 3 | 3 | – | – | – | – | – | – | 3 | 3 | 3 | 3 | 3 | 3 |
| Balls | 3 | 3 | – | – | – | – | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Dinosaur | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | – | – | 3 | 3 | 3 | 3 |
| Armadillo | 3 | 3 | 3 | 3 | 3 | 3 | – | – | – | – | 3 | 3 | 3 | 3 |

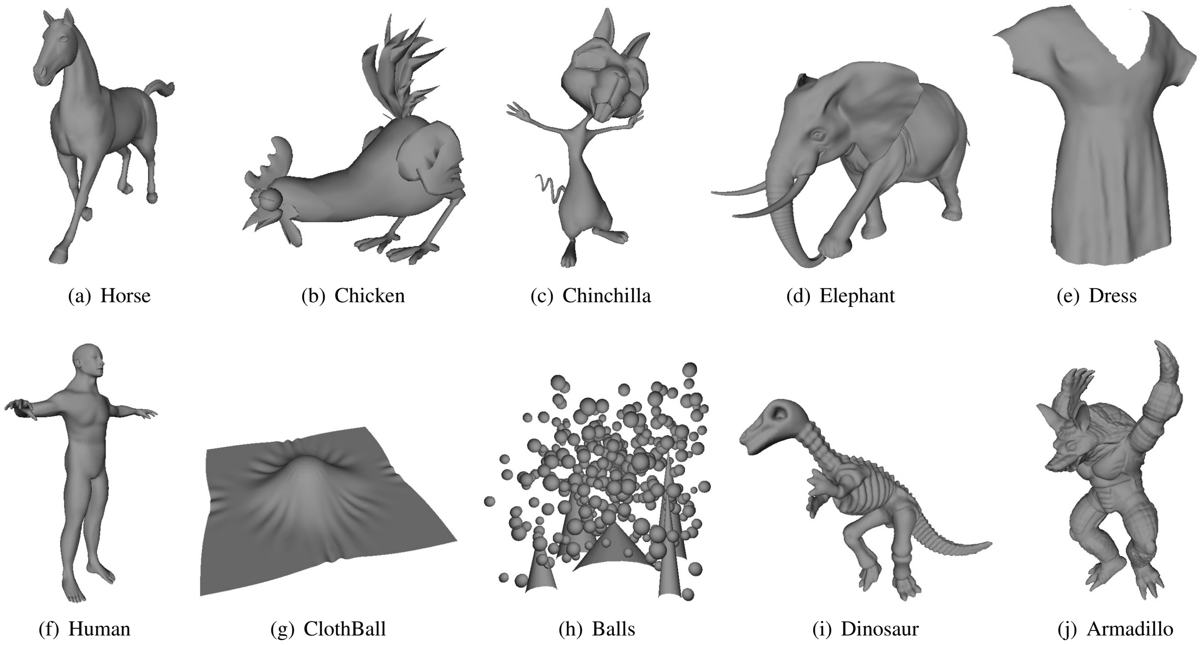
### *3.4. Distortions*

The purpose of our experiments is to collect ground-truth subjective measures of the perceptual quality of impaired dynamic meshes. We have included in our database a large number of 276 distorted mesh sequences. Table 2 summarizes the included distortions. For each distortion type, 3 intensity levels were generated for each selected mesh sequence, to cover a range of *high*, *medium* and *low* quality of the distorted meshes, which correspond respectively to low, medium and high distortion intensities. Distortion intensities were fixed through a small-scale preliminary study with few observers before the real test sessions, so as to achieve consistency across models and distortion types. This consistency aims to achieve roughly the same perceptual quality for animations impaired by the same level of intensity (low, medium or high) and has also been taken into account in the construction of popular image and video perceptual quality databases [18, 19, 20]. Considered distortions are classified into two main categories: simulated distortions and real-world distortions.

#### *3.4.1. Simulated distortions*

Simulated distortions include the so-called global distortions and descriptor-weighted distortions.

**Global distortion.** To apply global distortions, a random noise value was generated for each vertex coordinate in each frame of the dynamic mesh. The generated random additive noise follows either the *uniform* or the *Gaussian* distribution. When adding noise we did not set a control criterion to prevent creating self-intersections, mainly because in real applications many "careless" distortions can also generate such cases. Random noise addition can introduce both spatial and temporal visual distortions and is interesting for studying the combined effect of impairments in shape geometry and motion. We chose to add noise in the $(x, y, z)$ vertex coordinates mainly for the sake of simplicity. However, this kind of noise is also perceptually relevant. Indeed, according to a recent study [31], there is approximately a linear relationship between the noise in the $(x, y, z)$ coordinates and the induced normal vector angle change of a



(a) Horse (b) Chicken (c) Chinchilla (d) Elephant (e) Dress

(f) Human (g) ClothBall (h) Balls (i) Dinosaur (j) Armadillo

Figure 3: Snapshots of one single frame of the 10 reference mesh animations included in the subjective database.

facet. The facet normal vector is actually an essential element for the rendering of a triangle mesh. We believe that in the future it is important to study the perceptual influence of different kinds of noises in different coordinate systems (*e.g.*, in the global $(x, y, z)$ system or in the local tangent-normal system), or even for a noise in an arbitrary direction when applied on a vertex.

**Descriptor-weighted distortion.** The purpose of including descriptor-weighted distortion is to simulate and study the spatial and temporal visual masking effects. Such subjective studies on the spatial and temporal visual masking effects are missing in the literature of dynamic mesh PQA, whereas these effects are of great importance in the quality assessment of 3D mesh animations. In this type of descriptor-weighted distortion, the amount of alteration applied on each vertex coordinate is proportional or inversely proportional to the roughness of local surface ($LR_{ij}$ in Eq. (1)) or to the vertex speed ($S_{ij}$ in Eq. (2)). Note that spatial visual masking distortions were applied only on four dynamic meshes with both smooth and rough regions: Dinosaur, Chicken, Armadillo and Human. We believe that such meshes facilitate the observation and study of spatial visual masking effect. For the same reason, to simulate the temporal visual masking effect, speed-weighted distortions were applied only on meshes with both slow and rapid motions (such motions can be either local or global): Chicken, Horse, Elephant, ClothBall, Balls and Dinosaur. Two different kinds of noises were implemented to generate descriptor-weighted distortions:

- Same noise value for a given vertex over all frames (columns $SM_1$ and $TM_1$ in Table 2): a uniformly distributed random noise sequence of size $n_v$ (number of vertices in each frame of the original mesh sequence) was generated as $u_j^{(x)}, j \in \{1, 2, ..., n_v\}$. The additive noise was then weighted by $w_{ij}$, a coefficient proportional, or inversely proportional, to the local roughness (for $SM_1$) or to the speed (for $TM_1$) estimated at $v_{ij}$. Distorted $x$ component of the vertices was obtained as $\hat{v}_{ij}^{(x)} = v_{ij}^{(x)} + w_{ij} \times u_j^{(x)}, i \in \left\{1, 2, ..., n_f\right\}, j \in \{1, 2, ..., n_v\}$. The procedure is similar for $y$ and $z$ components.

- Same noise value for all vertices in a frame (columns $SM_2$ and $TM_2$ in Table 2): a uniformly distributed random noise sequence of size $n_f$ (number of frames in the mesh sequence) was generated as $u_i^{(x)}, i \in \left\{1, 2, ..., n_f\right\}$. The noise was then weighted by $w_{ij}$, a coefficient proportional, or inversely proportional, to the local roughness (for $SM_2$) or to the speed (for $TM_2$) at $v_{ij}$. As an example, the distorted $x$ component was obtained as $\hat{v}_{ij}^{(x)} = v_{ij}^{(x)} + w_{ij} \times u_i^{(x)}, i \in \left\{1, 2, ..., n_f\right\}, j \in \{1, 2, ..., n_v\}$. The procedure is similar for $y$ and $z$ components.

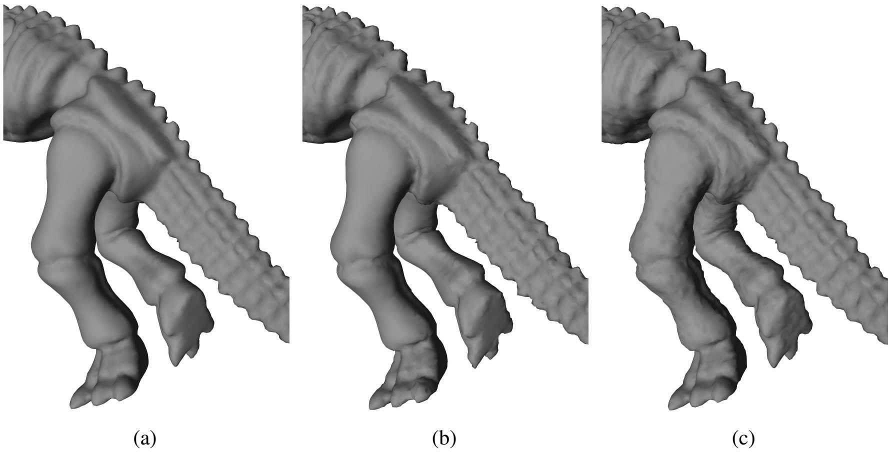


Figure 4: Visual effect of the spatial visual masking distortion $SM_1$ applied on Dinosaur (close-up images). (a)- Original frame, (b)- Type I noise weighted by roughness, and (c)- Type I noise inversely weighted by roughness. The noise is of medium intensity. Please refer to the electronic version of the article for a better visibility.

The first kind of noise (hereafter called Type I noise) introduces nearly pure spatial distortion, since the noise pattern is the same for all the frames. The second kind of noise (hereafter called Type II noise) introduces nearly pure temporal distortion, *i.e.*, shifting/trembling effect between frames. By weighting (or inversely weighting) Type I and II noises using local roughness in $SM_1$ and $SM_2$, and using vertex speed in $TM_1$ and $TM_2$, our objective is to study the capability of rough surface and fast motion in concealing spatial and temporal noises. It is interesting to point out the visual effect of each type of visual masking distortion. $SM_1$ distortion introduces a high-frequency spatial noise weighted or inversely weighted by local roughness (*WR* and *IWR* in Table 2); in general, distortions appear to be more visible in smooth regions than in rough regions. $SM_2$ distortion causes a translation/shifting of clusters of vertices having similar roughness, creating border effects between clusters of rough or smooth regions over successive frames. In $TM_1$ distortion, we add spatial high-frequency noise whose amplitude is weighted or inversely weighted by vertex speed (*WS* and *IWS* in Table 2); it is expected that the motion of fast regions can hide, to some extent, the added spatial noise. $TM_2$ noise introduces shifting/trembling effect between frames in either fast or slow regions; it appears that fast motion can effectively mask the visibility of this temporal distortion. Figure 4 illustrates an example of spatial visual masking simulation in $SM_1$, through a single frame of the Dinosaur sequence. As expected, the Type I noise weighted by roughness in Fig. 4.(b) is less visible than the same noise inversely weighted by roughness in Fig. 4.(c). For temporal effect of the applied distortions, readers could refer to videos on the database website at `http://www.gipsa-lab.fr/~kai.wang/software/database/`.

### *3.4.2. Real-world distortions*

Including real-world distortions is relevant when constructing a subjective database. In our experiments, realistic distortions were derived from lossy compression algorithms and network transmission error.

**Compression.** Among many existing dynamic mesh compression methods [17, 32, 33], two popular algorithms were selected: the *Motion Picture Experts Group* (MPEG-4) standardized algorithm of *Frame-based Animated Mesh Compression* (FAMC) [34] and the *COnnectivity Driven DYnAmic mesh Compression* algorithm (CODDYAC) [35].

The first compression algorithm (FAMC) is based on decomposing mesh frames into clusters and representing motions by a set of affine transformations. Visual distortions can occur at the border of each cluster. In order to reduce this border effect, a motion compensation step is carried out in the compression algorithm to smooth the predicted motion of vertices by combining the transformation of their proper cluster with those of the neighboring clusters. To generate compressed meshes, we used two different modes to encode prediction residues: the Discrete Cosine Transform (*DCT*) and the lifting bi-orthogonal wavelet transform (*Lifting*). For each mode, we generated three appropriate compression levels for each dynamic mesh by varying quantization levels of prediction residuals in the range of 6 to 9 bits. FAMC introduces spatial high-frequency granular distortions on mesh surfaces, coupled with motion compensation errors characterized by blockiness effect that appears between neighboring clusters. In general, granular distortion is more visible with DCT, while blockiness effect is more visible with lifting wavelet transform.

The second dynamic mesh compressor included in our study (CODDYAC) makes use of the temporal coherence of mesh vertices to carry out compression. A temporal PCA (Principal Component Analysis) representation is used to encode vertex trajectories, combined with a lossless mesh connectivity compressor. To generate compressed meshes, three parameters were varied: the number of encoded PCA basis vectors $N_b$, the quantization bits of PCA basis $Q_{pca}$, and the quantization bits of residual vectors $Q_r$. Considered ranges for $N_b$, $Q_{pca}$ and $Q_r$ were respectively $[30, 120]$, $[10.5, 16.6]$ and $[0.3, 2.7]$. Three compression levels were generated through different combinations of the three parameters. Reducing the number of PCA basis vectors and its quantization bits introduces a motion compensation mismatch characterized by sinusoidal-like impairments on the animation surface, while reducing the quantization bits of residual vectors leads to spatial high-frequency distortions.

**Network transmission error.** Considering network transmission errors in the scope of 3D mesh PQA is important and new in the literature. Indeed, with the consistent bandwidth increase of network infrastructure, 3D dynamic meshes are now more and more transmitted on IP networks. IP networks provide a best-effort delivery service. Despite this fact, transmission error may occur and it is important to study the effect of this error on the perceptual quality impairment of 3D dynamic meshes. To this end, dynamic meshes were firstly encoded with very high quality (nearly lossless) using FAMC before transmission. Meshes were encoded in *Group Of Pictures* (GOP) structure, and each GOP comprised 16 frames. The selected GOP pattern is $\{IBBBPBBBPBBBPBBB\}$, which contains 3 predicted frames $P$ between two $I$ frames (the two $I$ frames are in two different GOPs) and 3 bi-predicted frames $B$ between two anchor frames ($I$ or $P$). Then, like in [20], network error patterns were generated using a two-state Gilbert model [36] for a chosen packet-loss rate. Packets of the encoded mesh were dropped according to the generated error pattern. We did not drop the first packet, since it contains the FAMC header, needed to decode the animation. Average error-burst length of the network was set to 3 packets, in accordance with the result of a previous study [37]. Three packet-loss rates were selected from the set of $\{1.5\%, 2\%, 5\%, 10\%, 20\%\}$, to generate three levels of perceptual quality impairment. In general, IP network transmission error results in visual distortion transient in time and localized in space. High packet-loss rate can also introduce an annoying jerky motion effect to the mesh animation.

## 4. Results of Subjective Experiments

### *4.1. Processing and analysis of subjective scores*

After collecting the raw subjective scores, statistical processing and analysis are carried out to derive mean opinion scores and to well understand the obtained results.

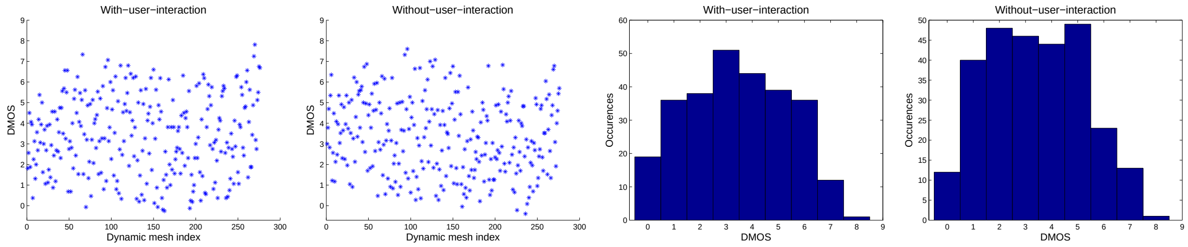
#### *4.1.1. Outlier rejection*

A screening procedure based on raw scores is applied in order to detect observers considered as outliers and remove their raw scores from the subsequent processing and analysis. Although test conditions are the same for all observers, there is commonly a variation in the judgment of perceptual quality between participants (systematic shifts when observers are too optimistic or too pessimistic, differences in using the quality scale, *etc.*). It is important to consider these aspects when detecting outliers. The outlier detection is carried out by using a well-established algorithm recommended by ITU [4]. We organized experiments in three sessions for either with- or without-user-interaction test condition. For with-user-interaction sessions, 4 outliers are detected among 75 participants. For without-user-interaction sessions, we detect only 1 outlier among 48 participants.

#### *4.1.2. Mean opinion scores*

After the outlier removal, differential scores $d_{ijk}$ are computed by subtracting the raw score $s_{ijk}$ of each distorted mesh sequence $i$ assigned by observer $j$ in session $k$ from the raw score $s_{rjk}$ of the corresponding reference sequence in the same session:

$$d_{ijk} = s_{rjk} - s_{ijk}. \tag{3}$$



(a) (b) (c) (d)

Figure 5: Scatter plots of $DMOS$ values of the impaired dynamic meshes for (a)- with-user-interaction experiments and (b)- without-user-interaction experiments, and the histograms of $DMOS$ values from (c)- with-user-interaction experiments and (d)- without-user-interaction experiments.

Finally, $d_{ijk}$ values are averaged over $N$ retained observers to compute a *Differential Mean Opinion Score* ($DMOS$) for each impaired dynamic mesh $i$ in test session $k$:

$$DMOS_{ik} = \frac{1}{N}\sum_{j=1}^{N} d_{ijk}. \tag{4}$$

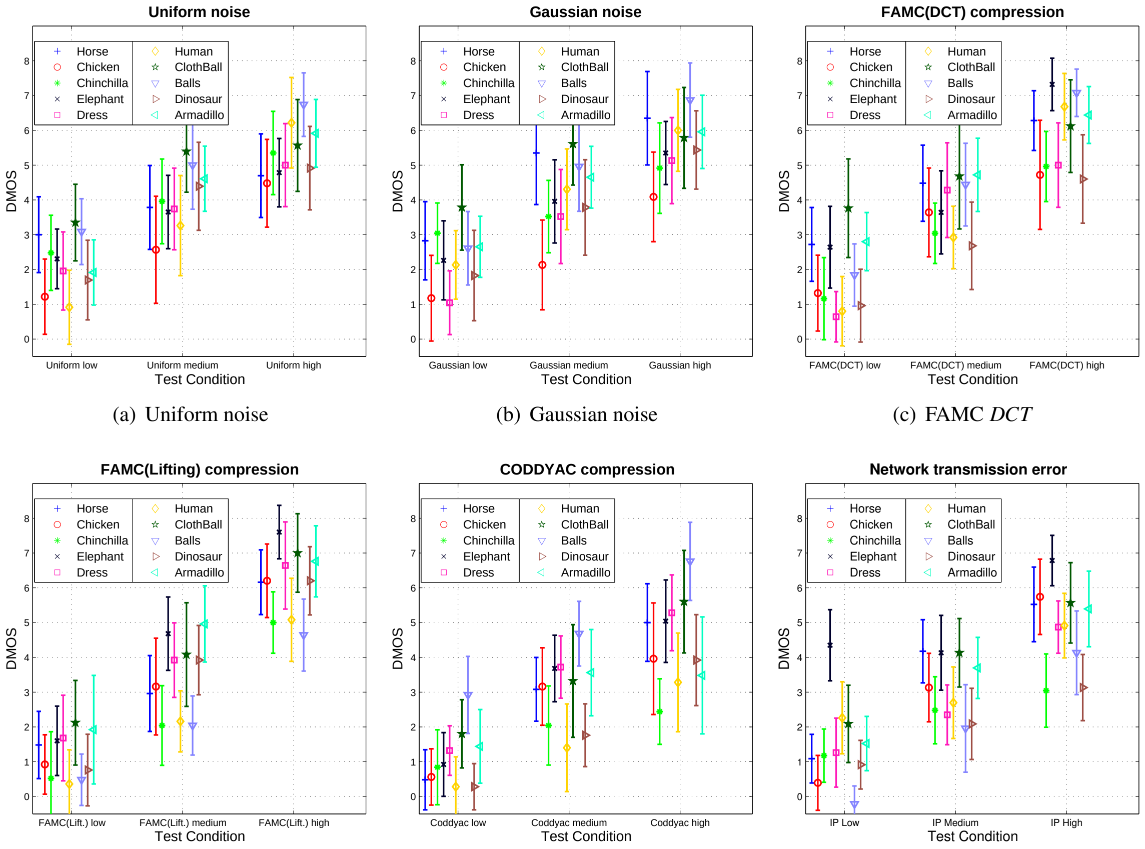
Figure 5 shows the scatter plots and the histograms of $DMOS$ values obtained from with-user-interaction and without-user-interaction experiments. In our subjective study observers rated each visual content in the range of 0 to 10. We find that the distribution of $DMOS$ values span about 80% of the available scale, which is comparable or even slightly larger than the covered ranges in existing video quality databases [19, 22, 38]. This demonstrates that the included mesh sequences cover a broad range of perceptual quality, which would be helpful for the evaluation of objective metrics. The processing based on differential scores aims to evaluate the relative opinion on the mesh perceptual quality, which is well suited to evaluate the performance of full-reference objective metrics. To evaluate no-reference objective metrics, it would be better to directly use *Mean Opinion Scores* ($MOS$). In that case, differential scores $d_{ijk}$ in Eq. (4) is replaced by raw scores $s_{ijk}$, to compute the $MOS$ values as:

$$MOS_{ik} = \frac{1}{N}\sum_{j=1}^{N} s_{ijk}. \tag{5}$$

### *4.1.3. Analysis of subjective scores*

We present $DMOS$ values associated with 95% confidence intervals in Fig. 6. Confidence intervals are computed using the *Student's t*-distribution. As shown in Fig. 6, for each distortion type, there are three distinct clusters of $DMOS$ values which correspond well to low, medium and high distortion levels introduced in the experiments. Few exceptions exist, for exemple the Elephant sequence distorted by low-level network transmission error is judged to have a slightly worse perceptual quality than Elephant impaired by medium-level network transmission error (see Fig. 6.(f)). We have re-examined the two sequences and have found that they are indeed of comparable perceptual quality, well reflected by the two $DMOS$ values. Low-level distortion introduces very localized spatial high-frequency distortion on Elephant legs that lasts for a number of frames, while medium-level distortion introduces both local (spatilly high-frequency artifacts) and global (shape contraction and expansion) distortions within a larger spatial area but more transient in time. Although we lost the oppotunity to include real low-level distortion of network transmission error on Elephant sequence, this mistake has very limited impact on the whole subjective database.

**Results of visual masking simulations**. In Fig. 7, we present average $DMOS$ values obtained from the four kinds of visual masking simulation distortions. For each distortion type, average $DMOS$ is computed over all involved sequences impaired by distortions of the same intensity level. For $SM_1$ distortion, average $DMOS$ values of the meshes on which the noise amplitude is inversely proportional to the local roughness are higher (therefore of worse perceptual quality) than those of the impaired meshes where the noise amplitude is proportional to the local roughness. In $SM_2$ distortion, a same random noise value was generated for all the vertices in each frame and weighted or



(a) Uniform noise

(b) Gaussian noise

(c) FAMC *DCT*

(d) FAMC *Lifting*

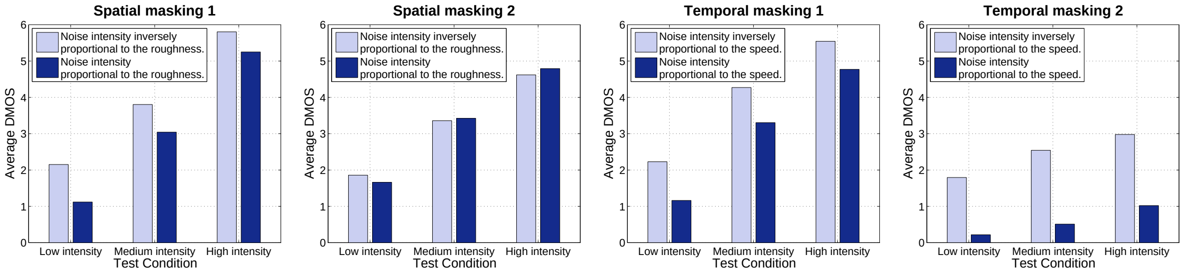
(e) CODDYAC

(f) Network transmission error

Figure 6: *DMOS* values and corresponding 95% confidence intervals for impaired models included in the subjective database. *DMOS* values from visual masking simulations are not plotted here and are analyzed in a different way (see the part of "Results of visual masking simulations").

inversely weighted by the local roughness. Applying this kind of noise leads to a temporal flickering effect in the spatial clusters having nearly the same roughness. As shown in Fig. 7.(b), there is almost no difference in average *DMOS* values for meshes impaired by additive noises weighted or inversely weighted by the local roughness in $SM_2$, since the cluster border and the flickering effect remain almost the same. Due to human visual system (HVS) limitations, empirically fast motion can reduce the visibility of distortions, and this visual masking effect has been utilized in perceptually-driven level-of-detail control methods of polygonal meshes [39]. This "temporal masking" phenomenon can be partially explained by the effect of temporal frequency on the shape of contrast sensitivity function (CSF) of HVS, where the sensitivity becomes very low when the temporal frequency is very high [40]. Across $TM_1$ and $TM_2$ distortions, average *DMOS* values of the meshes where noise amplitude is proportional to the vertex speed are always lower than those of the sequences where noise amplitude is inversely proportional to the speed (see Figs. 7.(c) and (d)). Translating/shifting vertices according to vertex speed is in general difficult to be observed. For that reason, average raw scores in $TM_2$ are generally higher (therefore average *DMOS* values are lower) than scores obtained under other distortion types.

To our knowledge, the subjective study on visual masking simulation presented here and the relevant analysis represent the first attempt in the literature of 3D dynamic mesh PQA. At this stage, two general observations that we can get from the results of visual masking simulation distortions are: 1) fast motion can conceal both spatial and temporal noises; and 2) rough surface can hide spatial noise, but is not capable of concealing temporal noise. We believe that in order for an objective metric to perform well in predicting dynamic mesh perceptual quality, it is important to take into account these spatial and temporal visual masking effects in the design of the metric.



(a) Spatial masking $SM_1$ (b) Spatial masking $SM_2$ (c) Temporal masking $TM_1$ (d) Temporal masking $TM_2$

Figure 7: Average $DMOS$ values for the four kinds of spatial and temporal visual masking simulation distortions. (a)- $SM_1$ distortion, (b)- $SM_2$ distortion, (c)- $TM_1$ distortion, and (d)- $TM_2$ distortion. For each type of distortion, $DMOS$ values are averaged over all involved meshes impaired by distortions of the same intensity level (low, medium or high) to obtain the average $DMOS$.

### 4.2. *Effect of user interaction*

In practical applications such as video games, we often allow the user to interact with the displayed 3D dynamic meshes, by using a mouse in most cases, or at least the camera position and orientation are changed during the display to let the observer get a full impression of the 3D object. Occasionally, dynamic meshes are displayed under a fixed viewpoint, as in the without-user-interaction experiments. A natural question is to know whether the user interaction has an effect on the perceived quality of dynamic meshes. In order to study this user interaction effect, we perform an analysis of $DMOS$ values obtained from with- and without-user-interaction experiments. More precisely, after verifying the Gaussianity of $DMOS$ values from the two test conditions using *Lilliefors* test, *paired-sample t-test* is carried out to examine the presence of "user interaction effect" on the resulted opinion scores. The null hypothesis of the statistical test can be considered equivalent to the statement that "Observer interaction does not statistically affect $DMOS$ values." For the distortion types where the interaction effect exists (*i.e.*, where the null hypothesis is rejected), we perform a further analysis: the percentage $\phi$ of the occurrences where $DMOS$ under with-interaction test is higher than its counterpart from without-interaction test is computed.

Results of t-tests are presented in Table 3, along with the $\phi$ values where the interaction effect exists. Paired-sample t-test rejects the null hypothesis for all distortion types except $SM_1$, $TM_2$ and FAMC. Observers are more severe on rating the quality of dynamic meshes impaired by uniform and Gaussian noises under with-interaction condition: 80.00% and 73.33% of the $DMOS$ values from with-interaction sessions are higher than those obtained under without-interaction test, for uniform and Gaussian noises, respectively. Such behavior can be explained by the fact that these two kinds of distortions are applied on the entire mesh, so with the possibility to zoom and rotate the mesh observers become more confident about the presence and intensity of the noise. On the other hand, observers are less severe in with-interaction sessions for $SM_2$, $TM_1$, CODDYAC and network error. $DMOS$ values from with-interaction sessions are in most cases below those from without-interaction sessions. The nature of these distortions helps us to understand the behavior of observers. All the four kinds of processing operations indroduce distortions that are more or less transient in time and/or localized in space. It is possible that the visibility of such distortions is decreased while enabling user interaction with the animation, which leads to uncertanity of the observer about the presence and intensity of the distortion. Finally, there is no statistically significant difference between $DMOS$ values for $SM_1$ and $TM_2$. One possible explanation is that the visual masking effect, quite obvious in $SM_1$ and $TM_2$, constitutes the most influential factor on the subjective scores, regardless of the test condition. We do not have convincing argument to explain the absence of user interaction effect under FAMC. One observation is that FAMC introduces distortions that are rather spatially and temporally consistent. This may make the distortion visually prominent, and easy to percieve and evaluate, under either with- or without-interaction condition.

The study conducted here shows that under most kinds of distortions user interaction can affect the perceived quality of 3D dynamic meshes. We are aware that a future in-depth investigation is necessary to understand the intrinsic mechanism of this influence, which is however beyond the scope of this paper. As mentioned earlier in Sections 1 and 3.1, another important motivation to conduct experiments under two different test conditions is to easily and accurately evaluate the performances of different kinds of objective metrics, *i.e.*, model-based metrics

Table 3: Analysis of the user interaction effect on $DMOS$ values: paired-sample t-test decision at 5% significance level and its corresponding $p$-value are provided. In the second column, '0' means that there is no user interaction effect, whereas '1' means that there is an interaction effect. $\phi$ indicates the percentage of occurrances where $DMOS$ from with-interaction test is higher than its counterpart from without-interaction test.

| Distortion | t-test decision | $p$-value | $\phi$ (%) |
|---|---|---|---|
| Uniform | 1 | 0.000 | 80.00 |
| Gaussian | 1 | 0.000 | 73.33 |
| $SM_1$ | 0 | 0.444 | – |
| $SM_2$ | 1 | 0.009 | 16.67 |
| $TM_1$ | 1 | 0.001 | 25.00 |
| $TM_2$ | 0 | 0.220 | – |
| FAMC | 0 | 0.806 | – |
| CODDYAC | 1 | 0.000 | 26.67 |
| Network error | 1 | 0.029 | 23.33 |

and image- and video-based metrics. In the next section, we will use the subjective scores collected from the with-user-interaction sessions to evaluate the performance of model-based metrics, as these metrics are designed to be independent of viewpoint used for the display of 3D meshes. For the evaluation of image- and video-based metrics, we will use the subjective scores collected from the without-user-interaction sessions, as such metrics are dependent on the selected viewpoint.

## 5. Performance Evaluation of Objective Metrics

With the subjective scores collected in our large-scale experiments, we will conduct in this section an evaluation and comparison of existing objective metrics for the task of 3D dynamic mesh PQA. Both model-based (Sections 5.1 and 5.2) and image- and video-based metrics (Section 5.3) are considered. It is worth mentioning the comprehensive aspect of our study on the performance evaluation of objective metrics, since we consider almost all the possible kinds of metrics that can be used for dynamic mesh PQA, including 2D (image-based) metrics, 2D+$t$ (video-based) metrics, 3D (model-based, designed for static meshes) metrics and 3D+$t$ (model-based, designed for dynamic meshes) metrics. To our knowledge, this is the first comprehensive comparative study of this type in the literature.

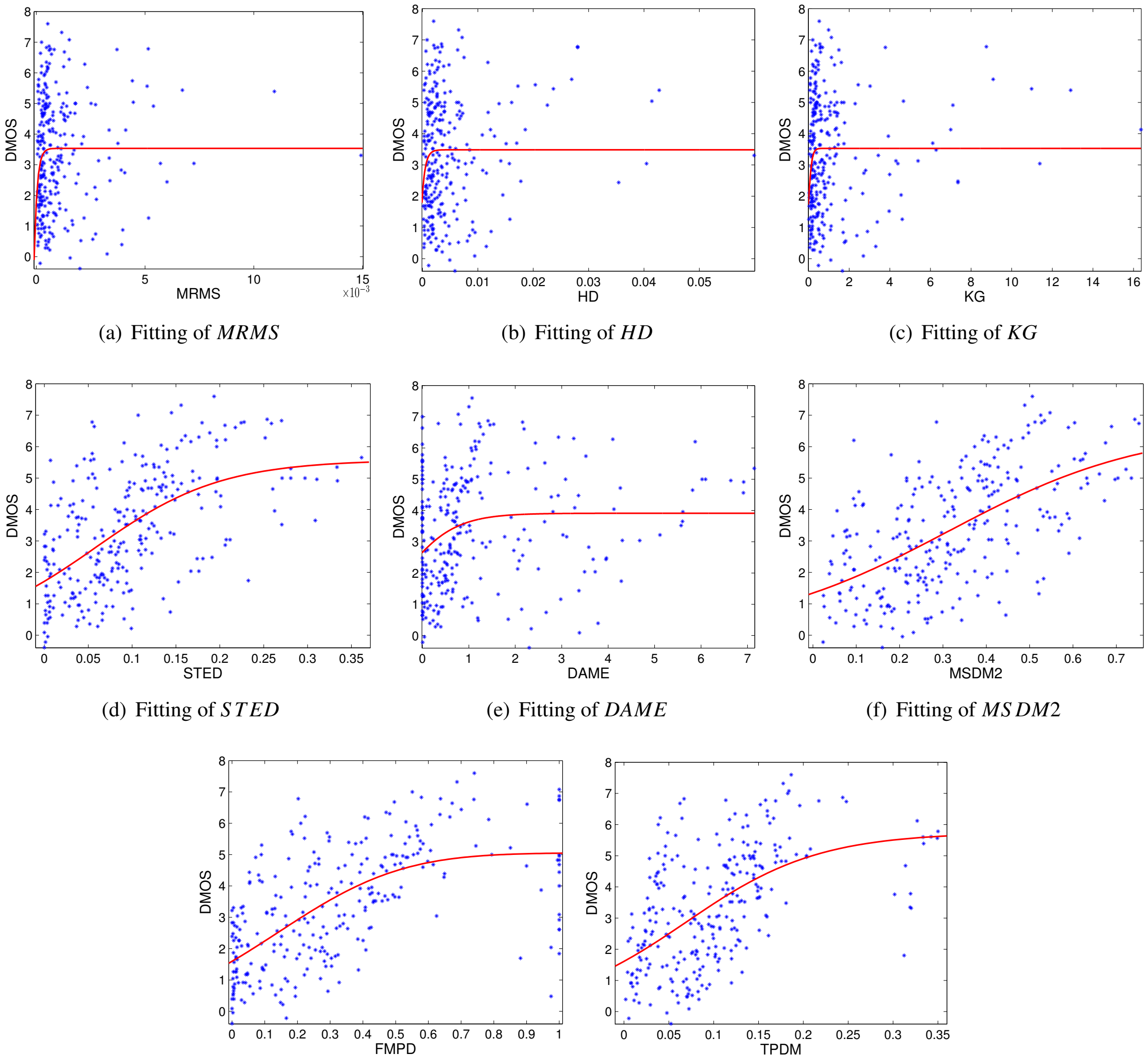
### 5.1. *Evaluation of model-based metrics*

We first evaluate the performance of recent model-based 3D mesh PQA metrics, on the constructed subjective database of dynamic mesh perceptual quality. The considered model-based objective metrics include:

- *Maximum Root Mean Square error* ($MRMS$) and *Hausdorff Distance* ($HD$) [41], two pure geometric distances between the surfaces of two static meshes.
- $KG$ metric [17], a pure geometric distance between two dynamic meshes based on vertex coordinate differences.
- $STED$ [16], a full-reference perceptually-driven 3D dynamic mesh PQA metric which is based on the comparison of mesh edge lengths and vertex displacements between two animations.
- Recent static mesh perceptually-driven quality metrics: $DAME$ [12], a full-reference metric based on the difference of dihedral angles of the two meshes under comparison; $MSDM2$ [13], a multi-scale full-reference metric that compares statistics of surface curvature amplitudes; $FMPD$ [14], a reduced-reference metric based on the comparison of surface global roughness; and $TPDM$ [15], a full-reference metric that uses both curvature amplitudes and curvature principal directions to compute a perceptual difference between two meshes.

For static mesh PQA metrics, per-frame quality measures are computed and averaged over all frames to obtain a global perceptual difference measure between two dynamic meshes.

In order to analyze and compare the performance of objective quality metrics, two criteria are used. The first criterion is the *Pearson Linear Correlation Coefficient* ($PLCC$), which measures the linear correlation between subjective and objective quality scores, *i.e.*, the prediction accuracy of the objective metrics. The second criterion is the *Spearman Rank Order Correlation Coefficient* ($SROCC$). It measures the correlation between the ranks of objective



(a) Fitting of *MRMS*

(b) Fitting of *HD*

(c) Fitting of *KG*

(d) Fitting of *STED*

(e) Fitting of *DAME*

(f) Fitting of *MSDM2*

(g) Fitting of *FMPD*

(h) Fitting of *TPDM*

Figure 8: Scatter plots of objective PQA measures (given by eight model-based metrics *MRMS*, *HD*, *KG*, *STED*, *DAME*, *MSDM2*, *FMPD* and *TPDM*) versus subjective *DMOS* values. The red curve is the fitted non-linear psychometric function.

and subjective scores, *i.e.*, the prediction monotonicity of the objective metrics. For a fair comparison, before computing the linear correlation *PLCC*, it is recommended to conduct a step of curve fitting between subjective scores and objective measures, using a chosen non-linear psychometric function [8, 18, 19]. *SROCC* is independent of the non-linear fitting if the used psychometric function is monotonic. In our study, we used the three-parameter logistic function, recommended by VQEG (Video Quality Experts Group) [38], for this fitting:

$$g(a, b, c, m) = \frac{a}{1 + \exp\left(-b\left(m - c\right)\right)}, \tag{6}$$

where $m$ is the original objective measure before fitting, and the parameters $a$, $b$ and $c$ are obtained through a non-linear least squares fitting (with the Matlab curve fitting toolbox), using the original objective scores and the corresponding *DMOS* values from the subjective database.

In Fig. 8, we show scatter plots of the objective quality predictions versus the ground-truth subjective scores, along with the fitted non-linear psychometric function. From these plots, we can observe a rather weak coherence between

Table 4: Pearson Linear Correlation Coefficient ($PLCC$, in %) after non-linear fitting of the model-based objective quality metrics. In bold are the highest $PLCC$ values provided by the best performing objective metric for each type of distortion (in column) or for the whole database (last column).

| Metric | Uniform | Gaussian | $SM_1$ | $SM_2$ | $TM_1$ | $TM_2$ | FAMC | CODDYAC | Network | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| $MRMS$ | 33.16 | 32.71 | 43.70 | 51.63 | 51.67 | −2.19 | 44.64 | 16.41 | 44.61 | 15.60 |
| $HD$ | 33.72 | 33.33 | 39.43 | 60.20 | 53.89 | −6.37 | 36.87 | 17.48 | 37.61 | 12.66 |
| $KG$ | 30.35 | 28.70 | 40.90 | 48.31 | 57.28 | −11.46 | 47.71 | 25.36 | 23.84 | 16.79 |
| $STED$ | 47.93 | 49.31 | 66.42 | 74.58 | **64.73** | 5.68 | 64.09 | 45.78 | 53.27 | 58.21 |
| $DAME$ | 12.13 | 14.39 | 43.01 | 36.83 | 23.35 | −5.39 | 17.62 | −5.37 | 34.77 | 23.76 |
| $MSDM2$ | 65.54 | 56.14 | **72.90** | **81.57** | 44.25 | −0.71 | 64.02 | 56.59 | **75.40** | 59.85 |
| $FMPD$ | **67.86** | **58.62** | 66.54 | 62.86 | 57.74 | **21.95** | **69.84** | **71.83** | 61.62 | **63.08** |
| $TPDM$ | 56.21 | 52.72 | 50.00 | 57.74 | 55.10 | 14.30 | 63.17 | 58.65 | 65.44 | 54.62 |

Table 5: Spearman Rank Order Correlation Coefficient ($SROCC$, in %) of the model-based objective quality metrics. In bold are the highest $SROCC$ values provided by the best performing objective metric for each type of distortion (in column) or for the whole database (last column).

| Metric | Uniform | Gaussian | $SM_1$ | $SM_2$ | $TM_1$ | $TM_2$ | FAMC | CODDYAC | Network | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| $MRMS$ | 32.06 | 29.60 | 31.57 | 64.54 | 45.33 | 9.83 | 50.13 | 37.91 | 68.03 | 11.46 |
| $HD$ | 32.33 | 31.38 | 29.05 | 57.75 | 41.42 | 2.78 | 43.64 | 38.40 | 69.81 | 11.37 |
| $KG$ | 32.24 | 27.55 | 27.96 | 65.19 | 45.59 | −0.52 | 51.99 | 40.67 | 58.93 | 11.54 |
| $STED$ | 45.46 | 47.25 | 67.10 | **85.98** | **68.57** | 12.53 | 65.52 | 44.37 | 62.22 | 57.66 |
| $DAME$ | 19.04 | 23.17 | 33.05 | 44.01 | 32.67 | −16.94 | 24.40 | 0.31 | 33.15 | 23.07 |
| $MSDM2$ | **66.48** | 52.17 | **72.19** | 80.41 | 40.98 | −3.61 | 66.59 | 57.40 | **79.43** | 59.58 |
| $FMPD$ | 63.59 | **57.29** | 67.67 | 61.58 | 54.00 | **25.32** | 63.60 | **71.25** | 68.08 | **62.81** |
| $TPDM$ | 59.67 | 55.06 | 48.23 | 58.32 | 56.56 | 13.75 | **66.78** | 65.64 | 69.12 | 55.61 |

the objective measures given by the eight model-based mesh PQA metrics and the $DMOS$ values of the subjective database. Tables 4 and 5 summarize the results of linear correlation and non-linear correlation, respectively, for each distortion type and for all the database ("Overall" columns in the tables). These results confirm the observation from Fig. 8, that is, objective measures do not correlate well with subjective scores. It appears that no objective metric can faithfully predict the perceptual quality of 3D dynamic meshes across all distortion types. Linear and non-linear correlations do not exceed 65% on the whole database (see "Overall" columns in Tables 4 and 5). In general, the simple geometric distances ($MRMS$, $HD$ and $KG$) provide the lowest correlation values with subjective scores, which again proves the need of developing perceptually-driven mesh PQA metrics. Surprisingly, compared with the perceptual quality metric $STED$ designed specifically for dynamic meshes, metrics designed for static meshes, in particular $MSDM2$ and $FMPD$, provide comparable, or even slightly better correlation results. $FMPD$ even has the highest $PLCC$ and $SROCC$ values on the whole database. Despite this fact, it seems inadequate to use static mesh PQA metrics to evaluate the perceptual quality of 3D dynamic meshes, because in general they do not have high correlation with subjective scores. The reason for the relatively low correlation is that static mesh PQA metrics do not take into account temporal distortions in the quality assessment.

Some metrics for 3D static meshes, especially $MSDM2$, explicitly account for the spatial visual masking effect, so they correlate relatively well under spatial visual masking distortion $SM_1$. They, as expected, in general fail to predict the perceptual quality in the presence of temporal visual masking effect. In particular, all the tested metrics provide poor correlation under $TM_2$ which simulates the pure temporal visual masking distortion. $STED$, the only existing perceptually-driven PQA metric for 3D dynamic meshes, does not show better performance. One possible explanation is that the temporal error in $STED$ has a small weight when combined with the spatial error [16]. Therefore, it is possible that the temporal distortion has not been sufficiently emphasized in the final score. Meanwhile, the spatial part of $STED$ has the potential to be further improved to better capture the spatial distortion [12]. In all, it appears that there is still some room for the performance improvement of model-based mesh PQA metrics. To this end, we may need to derive features that can successfully capture both spatial and temporal distortions, as well as the joint effect of these two kinds of distortions.

Table 6: Statistical comparison results of model-based metrics using prediction residual variances with 95% confidence criterion. Symbol '1' means that the metric in row is statistically better than the metric in column, while '0' means that the metric in row is statistically worse. Symbol '–' means that the performances of the pair of row/column metrics are statistically indistinguishable. Each symbol in a codeword indicates the statistical comparison result for a specific type of distortion, or the result on the whole database. The order of symbols is defined as follows: uniform noise, Gaussian noise, $SM_1$, $SM_2$, $TM_1$, $TM_2$, FAMC, CODDYAC, network transmission error, and overall (the whole database, in blod).

| | $KG$ | $STED$ | $DAME$ | $MSDM2$ | $FMPD$ | $TPDM$ |
|---|---|---|---|---|---|---|
| $KG$ | –––––––––**–** | 00000–000**0** | 110011110**0** | 000001000**0** | 00000–000**0** | 000001000**0** |
| $STED$ | 11111–111**1** | –––––––––**–** | 111111111**1** | 000011000**0** | 00011–000**0** | 001111–00**1** |
| $DAME$ | 001100001**1** | 000000000**0** | –––––––––**–** | 00000–000**0** | 000000000**0** | 000000000**0** |
| $MSDM2$ | 111110111**1** | 111100111**1** | 11111–111**1** | –––––––––**–** | 001100001**0** | 111100101**1** |
| $FMPD$ | 11111–111**1** | 11100–111**1** | 111111111**1** | 110011110**1** | –––––––––**–** | 111111110**1** |
| $TPDM$ | 111110111**1** | 110000–11**0** | 111111111**1** | 000011010**0** | 000000001**0** | –––––––––**–** |

### 5.2. Statistical comparison of model-based metrics

Statistical test is a rigorous way to faithfully assess the performance superiority (or inferiority, or equivalence) of an objective PQA metric over another. As pointed out by Sheikh *et al.* [18], the reliability of this statistical test increases as the number of impaired contents in the subjective database increases. Hence, as mentioned in Section 2, it is reasonable and safe to conduct such statistical tests on our large-scale subjective database as it includes a large number of impaired dynamic meshes. Similar to previous work on statistical comparison of image and video PQA metrics [18, 19, 21], a statistical test is performed in which we compare the variances of the residuals between the objective quality measures of model-based metrics (after non-linear fitting) and the $DMOS$ values for each distortion type and on the whole database.

Before conducting statistical tests, we tested the Gaussianity of the residual values using *Lilliefors* test. We have found that the residuals are not always Gaussian. Therefore, instead of *F-test*, we used *Levene's* test to compare the variances of the residues from two chosen objective metrics. Results of this statistical analysis are presented in Table 6. As shown in Tables 4 and 5, simple geometric distances ($MRMS$, $HD$ and $KG$) provide the lowest and similar correlations when tested on the subjective database. For the sake of simplicity, here in Table 6, we only consider $KG$ as the only pure geometric distance to be compared with the other metrics. From these statistical test results, we can clearly distinguish metric performances according to the variance comparison of their prediction residues. The overall performance of the metrics can be presented as the following ascending order: $KG$, $DAME$, $TPDM$, $STED$, $MSDM2$ and $FMPD$. Meanwhile, we have the following observations: $FMPD$ has statistically the best overall performance, as well as the best performance under uniform noise, Gaussian noise, FAMC and CODDYAC; $STED$ outperforms all the other metrics for $TM_1$; $MSDM2$ is the best metric to evaluate the perceptual quality of dynamic meshes impaired by network transmission error and by spatial visual masking distortions (both $SM_1$ and $SM_2$).

### 5.3. Evaluation of image- and video-based metrics

After the model-based metrics, we now study the performance of image- and video-based metrics when they are used for dynamic mesh PQA. The compared metrics include:

- The full-reference non-perceptual metric *Peak Signal to Noise Ratio* ($PSNR$).
- Two full-reference, 2D still-image perceptual quality metrics: the *Multi-Scale Structural SIMilarity* index ($MSSSIM$) [42] and the *Visual Information Fidelity* ($VIF$) measure [43].
- The recent no-reference, 2D still-image perceptual quality metric named *Naturalness Image Quality Evaluator* ($NIQE$) [44].
- Two video PQA metrics: the full-reference *Video Quality Metric* ($VQM$) [45] and the reduced-reference metric named *Spatio-Temporal Reduced Reference Entropic Differencing* ($STRRED$) measure [46].

The tested perceptual metrics have good performance in predicting image and video perceptual quality, as shown in the respective papers. We test $NIQE$ metric under two different settings for the training stage: 1) by using the 125 natural images given by the metric authors ($NIQE1$), and 2) by using 125 images of 2D snapshots of 3D meshes ($NIQE2$). Mesh snapshots are captured in the same software used for the subjective experiments, under the same

Table 7: Spearman Rank Order Correlation Coefficient ($SROCC$, in %) of the image- and video-based objective quality metrics when used for 3D dynamic mesh PQA. In bold are the highest $SROCC$ values provided by the best performing objective metric for each type of distortion (in column) or for the whole database (last column).

| Metric | Uniform | Gaussian | $SM_1$ | $SM_2$ | $TM_1$ | $TM_2$ | FAMC | CODDYAC | Network | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| $PSNR$ | 25.90 | 21.37 | 33.66 | 83.06 | 26.67 | −13.79 | 58.13 | 45.66 | 73.91 | 16.53 |
| $MSSSIM$ | 35.60 | 28.56 | 31.35 | 83.89 | 48.51 | 2.74 | 66.82 | 53.35 | 71.62 | 20.56 |
| $VIF$ | **35.78** | **33.85** | 42.53 | 87.02 | 46.07 | −13.31 | **67.98** | 55.60 | **74.67** | **27.39** |
| $NIQE1$ | 14.68 | 13.40 | −21.79 | −25.96 | −20.23 | **24.01** | 4.40 | 33.42 | 2.20 | 1.82 |
| $NIQE2$ | 17.20 | 14.62 | **43.62** | 23.70 | −18.10 | 13.70 | 7.34 | 28.23 | 8.28 | 12.68 |
| $VQM$ | 28.55 | 32.11 | 39.92 | **95.54** | **52.69** | −14.96 | 67.01 | 53.13 | 66.10 | 27.09 |
| $STRRED$ | 24.07 | 15.74 | −19.48 | −35.49 | −13.66 | 17.31 | 62.70 | **57.07** | 16.47 | 9.17 |

rendering condition. Meshes included in the training of $NIQE2$ are not in the subjective database but are similar to the reference meshes in the database. Table 7 presents the results of $SROCC$ values between the objective measures given by the evaluated image- and video-based metrics and the subjective scores collected from without-user-interaction sessions. Correlation values of the no-reference metrics $NIQE1$ and $NIQE2$ are computed using $MOS$ values, while $DMOS$ values are used to evaluate the performance of all the other metrics.

In general, image- and video-based metrics fail to correctly predict dynamic mesh perceptual quality, even under a fixed viewpoint as in the without-user-interaction experiments. The objective quality measures correlate poorly with subjective scores, as shown in Table 7. $VIF$, $VQM$ and $MSSSIM$, and even the non-perceptual metric $PSNR$, have relatively high correlation values under $SM_2$, FAMC and network transmission error distortions, but the overall correlation (see the last column of Table 7) remains quite low, being less than 30%. Training the no-reference metric $NIQE$ with 2D snapshots of 3D meshes only slightly improves its overall performance (see Table 7, compare the results of $NIQE1$ and $NIQE2$). Another observation is that video quality metrics $VQM$ and $STRRED$ do not perform better than image quality metrics such as $VIF$. In conclusion, it appears inadequate to use current image and video PQA metrics to evaluate the perceptual quality of 3D dynamic meshes, even under a fixed viewpoint. One possible explanation is that features used in current image- and video-based metrics are designed and optimized for PQA of *natural* images and videos, which are very different from *graphics* images and videos (such as mesh animation videos). A closely-related observation was reported in [47], where the authors showed that in general it was not adequate to use image quality metrics to conduct perceptual quality assessment of 3D geometric objects. Therefore, in the future, for the development of effective image- or video-based metrics for 3D mesh PQA, it would be important to understand the particularity of graphics images/videos when compared with natural images/videos. The goal is to derive specific 2D or 2D+$t$ features which can cope with the particular problem of static or dynamic mesh PQA.

## 6. A New Full-Reference Metric for 3D Dynamic Mesh PQA

In Section 5, we have shown that $FMPD$, although initially designed for static mesh PQA, has comparable, or even slightly better performance of dynamic mesh PQA than $STED$, a perceptual metric designed specifically for 3D dynamic meshes. $FMPD$ is a roughness based metric, and its basic assumption is that static mesh perceptual quality is closely related to the local and global surface roughness. In this section, we will first of all take local surface roughness as the spatial feature to construct a preliminary version of a full-reference model-based dynamic mesh PQA metric. Then, we will show that a simple vertex speed based weighting of roughness-based local perceptual error will lead to considerable improvement in terms of perceptual quality prediction performance, when tested on the constructed subjective database. Finally, the injection of two additional temporal features, respectively related to vertex speed and vertex moving direction, will further increase the correlation between objective measures and subjective scores.

### 6.1. Spatial-based perceptual distortion measure

We first present the preliminary version of the new metric where only the spatial feature is involved. Recall that our objective is to compute the perceptual distance between two dynamic meshes $\mathcal{M}_r$ and $\mathcal{M}_d$, $\mathcal{M}_r$ being a reference mesh and $\mathcal{M}_d$ being a distorted mesh. Let us still denote $n_v$ as the number of vertices in each frame of $\mathcal{M}_r$ (or $\mathcal{M}_d$), and $n_f$ as the number of frames in $\mathcal{M}_r$ (or $\mathcal{M}_d$). The corresponding vertex in $\mathcal{M}_d$ of a vertex $v^r_{ij}$ in $\mathcal{M}_r$ (*i.e.*, $j$-th vertex in $i$-th frame of $\mathcal{M}_r$) is denoted by $v^d_{ij}$. We first compute the local perceptual distance between vertices $v^r_{ij}$ and $v^d_{ij}$ as:

$$e^s_{ij} = \frac{\left|LR^r_{ij} - LR^d_{ij}\right|}{LR^r_{ij} + LR^d_{ij} + C_{LR}}, \tag{7}$$

where $LR^r_{ij}$ and $LR^d_{ij}$ are respectively the local roughness (non-negative values) at $v^r_{ij}$ and $v^d_{ij}$, and $C_{LR} = 0.002$ is a small constant to avoid numerical instability of the denominator near zero. Including a small constant in the denominator is a common practice in deriving an objective perceptual quality metric, *e.g.*, this is used in the well-known image Structural SIMilarity ($SSIM$) metric proposed in [48]. The local roughness is defined as the absolute value of the Laplacian of Gaussian curvature (see [14] for details). The local perceptual distance $e^s_{ij}$ can be viewed as a Michelson-like contrast definition in which $LR^r_{ij}$ and $LR^d_{ij}$ are involved. With this definition, $e^s_{ij}$ can, to some extent, capture the spatial visual masking effect. The reason is that a same local roughness change (*i.e.*, a same value of $\left|LR^r_{ij} - LR^d_{ij}\right|$) in a rougher region of the reference mesh (*i.e.*, where $LR^r_{ij}$ has a larger value) will lead to a smaller value of $e^s_{ij}$ , due to the term $LR^r_{ij}$ in the denominator of Eq. (7).

Afterwards, the local perceptual distances $e^s_{ij}, j = 1, 2, ..., n_v$ are spatially pooled together through a simple *Minkowski sum* to obtain the per-frame perceptual distance as:

$$f^s_i = \left(\sum_{j=1}^{n_v} \frac{1}{n_v} \left|e^s_{ij}\right|^{m_s}\right)^{\frac{1}{m_s}}, \tag{8}$$

with $i \in \left\{1, 2, ..., n_f\right\}$ the index of the frame, and we set the power parameter $m_s = 3$.

Finally, the per-frame distances are also temporally pooled together to obtain the purely spatial-based perceptual distance ($SPD$) between $\mathcal{M}_r$ and $\mathcal{M}_d$, as:

$$SPD_{\mathcal{M}_r,\mathcal{M}_d} = \left(\sum_{i=1}^{n_f} \frac{1}{n_f} \left|f^s_i\right|^{m_t}\right)^{\frac{1}{m_t}}, \tag{9}$$

with the power parameter $m_t = 4$. We have tried different combinations of the parameter values $m_s \in \{0.5, 1, 2, 3, 4\}$ and $m_t \in \{0.5, 1, 2, 3, 4\}$, and find that similar results are obtained if $m_s$ or $m_t$ is not too small, *i.e.*, $m_s \geq 2$ and $m_t \geq 3$. Such parameter settings will let the per-frame errors with higher values (and to some extent the local perceptual errors $e^s_{ij}$ of higher values) have higher impact on the final objective score. This is reasonable because empirically the subjective quality evaluated by human beings is more dependent on the highly impaired part (either spatial part or temporal part) of the visual content than on the high-quality/intact part, a widely accepted observation among the multimedia PQA research community [2]. At the end, we have empirically chosen $m_s = 3$ and $m_t = 4$ which give reasonable results when evaluated on the subjective database. Experimental results show that the spatial-based metric $SPD$ leads to slightly higher Spearman correlation value on the whole subjective database than all the existing metrics (see the second row of Table 8, and compare the results with those in Table 5). In particular, this simple purely spatial-based metric has very high correlation under $SM_1$, a kind of purely spatial visual masking distortion. This actually constitutes a good starting point for its further improvement to derive a well performing PQA metric which can also correctly evaluate the perceptual quality of distorted dynamic meshes impaired by some difficult distortions, where there is joint effect of spatial and temporal artifacts.

### *6.2. Speed-weighted spatial-based perceptual distortion measure*

In Section 5, we have shown that the visibility of spatial distortions can, to some extent, be reduced in fast moving regions of the dynamic mesh. We consider that in order to achieve high performance of dynamic mesh PQA, it is important to capture such hybrid spatio-temporal effect of perceived distortion. This motivates us to include a speed-based weighting of the local perceptual distance $e^s_{ij}$. The idea is very simple: $e^s_{ij}$ will be weighted by a small weight if the vertex $v^r_{ij}$ moves very fast, because as stated before, it would be more difficult to perceive the spatial distortion

Table 8: Spearman Rank Order Correlation Coefficient ($SROCC$, in %), on the constructed subjective database, of different versions of the proposed model-based metric for 3D dynamic mesh PQA.

| Metric | Uniform | Gaussian | $SM_1$ | $SM_2$ | $TM_1$ | $TM_2$ | FAMC | CODDYAC | Network | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| $SPD$ | 61.32 | 66.21 | 88.02 | 82.19 | 52.47 | 4.57 | 64.74 | 72.42 | 64.96 | 64.31 |
| $SWSPD$ | 72.33 | 77.88 | 91.72 | 87.45 | 68.52 | 1.44 | 68.78 | 85.53 | 82.21 | 72.91 |
| $DMPD$ | 83.39 | 76.85 | 91.54 | 96.67 | 70.05 | 93.52 | 71.26 | 90.11 | 88.80 | 80.14 |

in rapidly moving objects; on the contrary, the weight will be big if the vertex $v^r_{ij}$ stays almost still or moves very slowly. For the purpose of speed-based weighting, we first compute the forward speed of vertex $v^r_{ij}$ and denote it by $SP^r_{ij}$. Vertices in the last frame do not have forward speed, therefore in that case we consider $SP^r_{ij}$ as the vertex backward speed instead. In order to achieve scale-invariance of the proposed metric, the motion vectors (so also the vertex speed) are all normalized by the scene bounding box diagonal length of the dynamic mesh. We propose the following simple piecewise linear model for the speed-based weighting of $e^s_{ij}$:

$$e^{st}_{ij} = \begin{cases} w_1^{(SP)}.e^s_{ij}, & \text{if } SP^r_{ij} \leq Th_1^{(SP)}, \\ \left(w_1^{(SP)} - \frac{w_1^{(SP)}-w_2^{(SP)}}{Th_2^{(SP)}-Th_1^{(SP)}}\left(SP^r_{ij} - Th_1^{(SP)}\right)\right).e^s_{ij}, & \text{if } Th_1^{(SP)} < SP^r_{ij} \leq Th_2^{(SP)}, \\ w_2^{(SP)}.e^s_{ij}, & \text{if } SP^r_{ij} > Th_2^{(SP)}, \end{cases} \tag{10}$$

where $e^{st}_{ij}$ is the weighted local perceptual distance, and we set the parameter values as $Th_1^{(SP)} = 0.01$, $Th_2^{(SP)} = 0.05$, $w_1^{(SP)} = 1$ and $w_2^{(SP)} = 0.3$. In practice, we have tried many different values for the parameters $Th_1^{(SP)}$, $Th_2^{(SP)}$, $w_1^{(SP)}$ and $w_2^{(SP)}$, and their combinations. It is found that different parameter value combinations always lead to improvement in terms of overall correlation with subjective scores for the objective metric. This also implies the appropriateness of such a speed-based weighting of local perceptual errors. It is worth mentioning that without deeper investigation and understanding of the psychophysical mechanism of the HVS, it is indeed a common practice to use empirical functions and/or empirical parameter values in the development of objective PQA metrics, *e.g.*, in [13, 14, 48, 49]. As mentioned at the end of this section, a more perceptually relevant weighting scheme would be possible, probably based on recent findings about the motion perception of HVS, and this constitutes one part of our future work. Nevertheless, from Eq. (10), we can see that the spatial-based local perceptual distance $e^s_{ij}$ is mitigated in fast-moving regions, whereas it remains nearly untouched in still or slowly-moving regions of the mesh sequence. This corresponds well to the practical observations from our subjective experiments (see Section 4.1.3).

The weighted local errors $e^{st}_{ij}$ are then spatially pooled together to yield per-frame errors, still using the power-3 Minkowski sum, with a similar formula as Eq. (8). For the pooling of per-frame errors, we find that in order to achieve a higher correlation value with subjective scores, it would be better to replace the Minkowski sum by a more sophisticated pooling strategy which depends on the per-frame distortion severity. More precisely, we first sort the per-frame errors $f^{st}_i, i = 1, 2, ..., n_f$ in ascending order. The sorted per-frame errors are denoted by $\tilde{f}^{st}_i, i = 1, 2, ..., n_f$, where we have $\tilde{f}^{st}_{i_1} \leq \tilde{f}^{st}_{i_2}$ for any $i_1 \leq i_2$. A group of $n_f$ weights are then determined as $\tilde{w}_i = (\frac{i}{n_f})^{pw}, i = 1, 2, ..., n_f$, with $pw$ a parameter that controls the increasing speed of the weights $\tilde{w}_i$ along with the increase of the index $i$. Finally, the pooling of the per-frame errors is carried out as:

$$SWSPD_{\mathcal{M}_r,\mathcal{M}_d} = \frac{\sum_{i=1}^{n_f} \tilde{w}_i.\tilde{f}^{st}_i}{\sum_{i=1}^{n_f} \tilde{w}_i}. \tag{11}$$

We call the obtained pooled error $SWSPD$ as the speed-weighted spatial-based perceptual distance between $\mathcal{M}_r$ and $\mathcal{M}_d$. It can be seen that a higher value of the parameter $pw$ will assign higher weights to the frames with severer distortion. This in practice improves the correlation between objective and subjective scores, in a manner more effective than the Minkowski sum. In this paper, we choose $pw = 10$. The dynamic mesh PQA performance of the metric $SWSPD$ is reported in the third row of Table 8. We can see that noticeable improvement is achieved for $SWSPD$ when compared with $SPD$, *e.g.*, under distortions such as CODDYAC and network transmission error. The overall correlation has also been increased. Although the correlation value under $TM_1$ is not that high (68.52%) for $SWSPD$, an improvement of about 16% has been obtained compared with $SPD$. These results demonstrate the utility of the proposed speed-based weighting scheme of the roughness-based spatial local perceptual error.

### *6.3. Incorporating temporal features and the new metric*

In the following, we will show that incorporating additional temporal features to the above speed-weighted metric will further improve its performance. Two simple temporal features are considered here, which are respectively related to vertex speed and vertex moving direction.

For a vertex $v_{ij}^r$ in the reference mesh $\mathcal{M}_r$ and its counterpart vertex $v_{ij}^d$ in the distorted mesh $\mathcal{M}_d$, we compute the forward motion vectors of these two vertices and denote them by $\vec{d}_{ij}^r$ and $\vec{d}_{ij}^d$, respectively. Note that the motion vectors are still normalized by the scene bounding box diagonal length to ensure scale-invariance. It is not possible to compute forward motion vector for vertices in the last frame of the dynamic mesh, so in that case the backward motion vector is used instead. A motion-vector-norm-related (*i.e.*, vertex-speed-related) contrast is computed as:

$$s_{ij}^s = \frac{\left| \|\vec{d}_{ij}^r\| - \|\vec{d}_{ij}^d\| \right|}{\|\vec{d}_{ij}^r\| + \|\vec{d}_{ij}^d\| + C_S}, \tag{12}$$

where we set $C_S = 0.002$ to avoid numerical instability, and we call $s_{ij}^s$ the speed-based local perceptual distance. Note the similarity between Eq. (12) and Eq. (7). It is therefore not difficult to understand that $s_{ij}^s$ can, to some extent, capture the temporal visual masking effect: a same speed change, *i.e.*, a same value of $\left| \|\vec{d}_{ij}^r\| - \|\vec{d}_{ij}^d\| \right|$, will induce a smaller error $s_{ij}^s$ in fast-moving part of the mesh, *i.e.*, where $\|\vec{d}_{ij}^r\|$ is higher, than in slowly-moving or still part of the mesh. This is in accordance with the observation that motion alternation in a fast-moving object is more difficult to perceive than that in a slowly-moving or still object, as for example reflected by the $DMOS$ values corresponding to the $TM_2$ distortion in our subjective experiments (see Section 4.1.3).

A natural complement of the vertex-speed-based contrast in Eq. (12) would be a similar contrast defined according to the vertex moving direction. To this end, we first compute for each vertex its backward and forward motion vectors, and the angle between these two vectors is calculated and hereafter denoted by $AG_{ij}^r \in [0, \pi)$ for vertex $v_{ij}^r$ (and denoted by $AG_{ij}^d \in [0, \pi)$ for vertex $v_{ij}^d$). Theoretically, it is not possible to obtain this angle value for vertices in the first and the last frames. To remedy this problem, the angle values of such vertices are directly copied from the counterpart vertex in the second frame (for vertices in the first frame), or from the counterpart vertex in the second last frame (for vertices in the last frame). A vertex-moving-direction-related contrast is defined as:

$$a_{ij}^s = \frac{\left| AG_{ij}^r - AG_{ij}^d \right|}{AG_{ij}^r + AG_{ij}^d + C_{AG}}, \tag{13}$$

with $C_{AG} = \frac{1}{32}\pi$ a small constant to avoid numerical instability. Still note the similarity between Eq. (13) and Eqs. (7) and (12), and note that similar to $s_{ij}^s$, the derivation of $a_{ij}^s$ also takes into account the temporal visual masking effect.

The local perceptual distances $s_{ij}^s$ and $a_{ij}^s$ are spatially and temporally pooled via Minkowski sum (both with spatial power 5 and temporal power 3) to yield two global perceptual distances $SPPD_{\mathcal{M}_r,\mathcal{M}_d}$ and $APD_{\mathcal{M}_r,\mathcal{M}_d}$, respectively standing for the speed-based perceptual distance and the angle-based perceptual distance between $\mathcal{M}_r$ and $\mathcal{M}_d$. Here, we use the simple Minkowski pooling for $s_{ij}^s$ and $a_{ij}^s$, because we find that for these two *temporal* features, the distortion severity based weighting proposed in the last subsection does not help in improving the metric performance. As mentioned later, a thorough understanding and study on the adaptive spatial and temporal pooling strategy for each kind of spatial and temporal feature is one important part of our future work. Our first observation is that for a temporal feature, it would be more favorable to emphasize on the high-valued spatial distortions than on the high-valued per-frame distortions; on the contrary, for a spatial feature, it would be necessary to emphasize on the per-frame distortions with high values than on the local spatial distortions with high values (*e.g.*, notice the power values used in Minkowski sum here for the two *temporal* features and those used in Section 6.2 for the roughness-based *spatial* feature).

Now we have three kinds of perceptual distances $SWSPD$, $SPPD$ and $APD$. We hereafter use a simple weighted mean square combination to merge the three distances. The final dynamic mesh perceptual distance (*DMPD*) is computed as (for the sake of brevity, we drop in Eq. (14) the subscript $\cdot_{\mathcal{M}_r,\mathcal{M}_d}$ for $SWSPD$, $SPPD$ and $APD$):

$$DMPD_{\mathcal{M}_r,\mathcal{M}_d} = \sqrt{w_1.SWSPD^2 + w_2.SPPD^2 + w_3.APD^2}, \tag{14}$$

with the three weights empirically set as $w_1 = 0.4$, $w_2 = 0.5$ and $w_3 = 0.1$. This is actually a very simple error combination strategy, and further improvement is possible, *e.g.*, using advanced fusion methods. The last row of Table 8 shows that after including two additional temporal features, the metric performance is further improved. The new metric $DMPD$ has a global correlation value higher than 80% with subjective scores. In particular, $DMPD$ has very high correlation under $SM_2$ and $TM_2$, where the temporal trembling artifact is the dominant distortion. This proves the effectiveness of the two temporal features incorporated in the proposed metric.

In all, the proposed metric is conceptually simple and similarity-transformation-invariant. Moreover, it achieves relatively high correlation with subjective scores and thus shows the potential to be a well-performing dynamic mesh PQA metric. However, we by no means claim that the proposed metric performs well under any type of distortion or that it is mature enough to be faithfully used in practical applications. The metric can be improved in several aspects. For example, a more sophisticated and adaptive spatial and temporal pooling strategy might be devised for each spatial or temporal feature. It would also be interesting to make use of the findings about motion perception mechanism in the human visual system [50, 51], *e.g.*, to optimize the speed-based contrast definition and to derive a better weighting scheme of the spatial-based local perceptual error.

## 7. Conclusions and Future Work

With the increasing use of surface animations in various practical applications, the perceptual quality assessment of 3D dynamic meshes has become an important research problem. One obstacle of the rapid advance of relevant research is the lack of large-scale ground-truth data of subjective dynamic mesh perceptual quality, on which objective metrics can be reliably evaluated and compared. As analyzed in Section 2, the only existing subjective database of dynamic meshes has a very limited number of distorted meshes and some limitations. In this paper, firstly we have designed and constructed a new subjectively-rated database of 3D dynamic meshes. Our database comprises a large set of 276 impaired meshes generated from 10 reference animations by applying various real-world and simulated distortions. The subjective scores collected from our large-scale experiments have been properly processed and analyzed. In addition, a group of carefully designed distortions have been applied, for the purpose of studying the spatial and temporal visual masking effects, which is to our knowledge the first attempt in the context of dynamic mesh PQA. The main finding is that high roughness is able to hide pure spatial noise but not temporal noise, whereas high speed can conceal both spatial and temporal noises. The constructed database, along with the associated $MOS$ and $DMOS$ values, is freely available on-line at `http://www.gipsa-lab.fr/~kai.wang/software/database/`. Based on this subjective database, we have conducted a comparative study of the latest model-based and image- and video-based objective metrics when they are used for assessing the perceptual quality of 3D dynamic meshes. This new and comprehensive comparative study considers almost all the possible types of objective metrics for dynamic mesh PQA, including 2D, 2D+*t*, 3D and 3D+*t* metrics. In general, model-based metrics outperform image- and video-based metrics, but the performance of both kinds of metrics still needs improvement to achieve faithful objective dynamic mesh PQA. The results of this objective study reflect that the research on dynamic mesh PQA is still in its very early stage, and are also expected to provide some insights for the future development of more effective metrics. Indeed, in the last part of this paper, we have proposed a simple dynamic mesh PQA metric which makes use of both spatial and temporal features. In particular, during the three-step design of the proposed metric, we show that the injection of perceptually relevant temporal features is effective in gradually improving the performance of the objective metric.

We plan to continue our work on both subjective and objective PQA of 3D dynamic meshes. Extension of the subjective database is possible. We can include more types of distortions (*e.g.*, smoothing, sharpening and distortions introduced by different compressors [17, 32, 33] that are more or less different), as well as multiply distorted meshes and impaired color and textured mesh sequences. It would be interesting to conduct an in-depth investigation to further study the spatio-temporal visual masking effects in the context of dynamic mesh PQA. Finally, based on some properly defined spatio-temporal features, we would like to develop more effective objective metrics. These metrics could be either full-reference metrics or no-reference metrics, and could be either model-based or video-based.

## Acknowledgment

We would like to thank owners of the reference meshes included in the subjective database, for granting us permission to post their meshes on-line. We are grateful to all the participants of our subjective study. We thank Dr. G.

Lavoué for his useful suggestions on the manuscript, as well as the anonymous reviewers for their helpful comments. This work is in part supported by the Rhône-Alpes Region through the ARC6 funded project PADME.

## References

[1] M. Botsch, L. Kobbelt, M. Pauly, P. Alliez, B. Lévy, Polygon Mesh Processing, AK Peters, 2010.

[2] Z. Wang, A.-C. Bovik, Modern Image Quality Assessment, Morgan & Claypool, 2006.

[3] M. Corsini, M. Larabi, G. Lavoué, O. Petřík, L. Váša, K. Wang, Perceptual metrics for static and dynamic triangle meshes, Comput. Graphics Forum 32 (1) (2013) 101–125.

[4] International Telecommunication Union, Rec. BT.500: Methodology for the Subjective Assessment of the Quality of Television Pictures (2012).

[5] International Telecommunication Union, Rec. P.910: Subjective Video Quality Assessment Methods for Multimedia Applications (2008).

[6] S. Winkler, Analysis of public image and video databases for quality assessment, IEEE J. Sel. Topics Signal Process. 6 (6) (2012) 616–625.

[7] B. Watson, A. Friedman, A. McGaffey, Measuring and predicting visual fidelity, in: Proc. of ACM Siggraph, 2001, pp. 213–220.

[8] M. Corsini, E. Drelie Gelasca, T. Ebrahimi, M. Barni, Watermarked 3-D mesh quality assessment, IEEE Trans. Multimedia 9 (2) (2007) 247–256.

[9] G. Lavoué, E. Drelie Gelasca, F. Dupont, A. Baskurt, T. Ebrahimi, Perceptually driven 3D distance metrics with application to watermarking, in: Proc. of SPIE Electronic Imaging, 2006, pp. 63120L.1–63120L.12.

[10] G. Lavoué, A local roughness measure for 3D meshes and its application to visual masking, ACM Trans. Appl. Percept. 5 (4) (2009) 21:1–21:23.

[11] S. Silva, B.-S. Santos, C. Ferreira, J. Madeira, A perceptual data repository for polygonal meshes, in: Proc. of Int. Conf. in Visualization, 2009, pp. 207–212.

[12] L. Váša, J. Rus, Dihedral angle mesh error: a fast perception correlated distortion measure for fixed connectivity triangle meshes, Comput. Graphics Forum 31 (5) (2012) 1715–1724.

[13] G. Lavoué, A multiscale metric for 3D mesh visual quality assessment, Comput. Graphics Forum 30 (5) (2011) 1427–1437.

[14] K. Wang, F. Torkhani, A. Montanvert, A fast roughness-based approach to the assessment of 3D mesh visual quality, Comput. & Graphics 36 (7) (2012) 808–818.

[15] F. Torkhani, K. Wang, J.-M. Chassery, A curvature tensor distance for mesh visual quality assessment, in: Proc. of Int. Conf. on Computer Vision and Graphics, 2012, pp. 253–263.

[16] L. Váša, V. Skala, A perception correlated comparison method for dynamic meshes, IEEE Trans. Vis. Comput. Graphics 17 (2) (2011) 220–230.

[17] Z. Karni, C. Gotsman, Compression of soft-body animation sequences, Comput. & Graphics 28 (1) (2004) 25–34.

[18] H.-R. Sheikh, M. Sabir, A.-C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3440–3451.

[19] K. Seshadrinathan, R. Soundararajan, A.-C. Bovik, L.-K. Cormack, Study of subjective and objective quality assessment of video, IEEE Trans. Image Process. 19 (6) (2010) 1427–1441.

[20] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, T. Ebrahimi, A H.264/AVC video database for the evaluation of quality metrics, in: IEEE Int. Conf. on Acoust., Speech, Signal Process., 2010, pp. 2430–2433.

[21] A.-K. Moorthy, K. Seshadrinathan, R. Soundararajan, A.-C. Bovik, Wireless video quality assessment: A study of subjective scores and objective algorithms, IEEE Trans. Circuits Syst. Video Technol. 20 (4) (2010) 587–599.

[22] A.-K. Moorthy, L.-K. Choi, A.-C. Bovik, G. de Veciana, Video quality assessment on mobile devices: Subjective, behavioral and objective studies, IEEE J. Sel. Topics Signal Process. 6 (6) (2012) 652–671.

[23] International Telecommunication Union, Rec. BT.1788: Methodology for the Subjective Assessment of Video Quality in Multimedia Applications (2007).

[24] International Telecommunication Union, Rec. BT.1082: Studies toward the Unification of Picture Assessment Methodology (1990).

[25] R.-W. Sumner, J. Popović, Deformation transfer for triangle meshes, ACM Trans. Graphics 23 (3) (2004) 399–405.

[26] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, T. Boubekeur, Markerless garment capture, ACM Trans. Graphics 27 (3) (2008) 99:1–99:10.

[27] S.-E. Yoon, S. Curtis, D. Manocha, Ray tracing dynamic scenes using selective restructuring, in: Proc. of Eurographics Symposium on Rendering, 2007, pp. 73–84.

[28] M. Tang, S. Curtis, S.-E. Yoon, D. Manocha, ICCD: Interactive continuous collision detection between deformable models using connectivity-based culling, IEEE Trans. Vis. Comput. Graphics 15 (2009) 544–557.

[29] I. Baran, J. Popović, Automatic rigging and animation of 3D characters, ACM Trans. Graphics. 26 (3) (2007) 72:1–72:8.

[30] H. Yu, S. Winkler, Image complexity and spatial information, in: Proc. of Int. Workshop on Quality of Multimedia Experience, 2013, pp. 135–157.

[31] Y. Yang, N. Peyerimhoff, I. Ivrissimtzis, Linear correlations between spatial and normal noise in triangle meshes, IEEE Trans. Vis. Comput. Graphics 19 (1) (2013) 45–55.

[32] E.-S. Jang, J.-D.-K. Kim, S.-Y. Jung, M.-J. Han, S.-O. Woo, S.-J. Lee, Interpolator data compression for MPEG-4 animation, IEEE Trans. Circuits Syst. Video Technol. 14 (7) (2004) 989–1008.

[33] K. Müller, A. Smolic, M. Kautzner, P. Eisert, T. Wiegand, Rate-distortion-optimized predictive compression of dynamic 3D mesh sequences, Signal Process.: Image Commun. 21 (9) (2006) 812–828.

[34] K. Mamou, T. Zaharia, F. Preteux, FAMC: The MPEG-4 standard for animated mesh compression, in: Proc. of IEEE Int. Conf. on Image Process., 2008, pp. 2676–2679.

[35] L. Váša, V. Skala, Coddyac: Connectivity driven dynamic mesh compression, in: Proc. of 3DTV Conference, 2007, pp. 1–4.

[36] E.-N. Gilbert, Capacity of a burst-noise channel, Bell System Technical Journal 39 (1960) 1253–1266.

[37] T. Chua, D.-C. Pheanis, QoS evaluation of sender-based loss-recovery techniques for VoIP, IEEE Netw. 20 (6) (2006) 14–22.

[38] Video Quality Experts Group, Final Report from the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment Phase I. (2000).

[39] D. Luebke, M. Reddy, J. Cohen, A. Varshney, B. Watson, R. Huebner, Level of Detail for 3D Graphics, Morgan Kaufmann, 2003.

[40] J.-G. Robson, Spatial and temporal contrast sensitivity functions of the visual system, J. Opt. Soc. Amer. 56 (1966) 1141–1142.

[41] P. Cignoni, C. Rocchini, R. Scopigno, Metro: measuring error on simplified surfaces, Comput. Graphics Forum 17 (2) (1998) 167–174.

[42] Z. Wang, E.-P. Simoncelli, A.-C. Bovik, Multiscale structural similarity for image quality assessment, in: IEEE Asilomar Conf. Signals, Syst. Comput., 2003, pp. 1398–1402.

[43] H.-R. Sheikh, A.-C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.

[44] A. Mittal, R. Soundararajan, A.-C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.

[45] M.-H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.

[46] R. Soundararajan, A.-C. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, IEEE Trans. Circuits Syst. Video Technol. 23 (4) (2013) 684–694.

[47] B.-E. Rogowitz, H.-E. Rushmeier, Are image quality metrics adequate to evaluate the quality of geometric objects, in: Proc. of SPIE Human Vision and Electronic Imaging, 2001, pp. 340–348.

[48] Z. Wang, A.-C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[49] Z. Wang, L. Lu, A.-C. Bovik, Video quality assessment based on structural distortion measurement, Signal Process.: Image Commun. 19 (2) (2004) 121–132.

[50] J.-H. Maunsell, D.-C. Van Essen, Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation, J. Neurophysiology 49 (5) (1983) 1127–1147.

[51] N.-C. Rust, V. Mante, E.-P. Simoncelli, J.-A. Movshon, How MT cells analyze the motion of visual patterns, Nature Neuroscience 9 (11) (2006) 1421–1431.