ROYAUME DU MAROC
Université Mohammed V جامعة محمد الخامس
- RABAT - - الرباط -
Faculté des sciences
كلية العلوم

CENTRE D'ETUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

N° d'ordre : 3371

THESE

En vue de l'obtention du : ***DOCTORAT***

Structure de Recherche : Laboratoire de Recherche en Informatique et Télécommunications
Discipline : Sciences de l'ingénieur
Spécialité : Informatique et télécommunications

Présentée et soutenue le 14/11/2020 par :

Ilyass ABOUELAZIZ

Évaluation sans référence de la qualité des maillages 3D : Exploitation de l'apprentissage profond et la saillance visuelle

JURY

Mohamed OUADOU	PES, Université Mohammed V-Rabat, Faculté des Sciences	Président
Mohammed EL HASSOUNI	PES, Université Mohammed V-Rabat, Faculté des Lettres et des Sciences Humaines	Directeur de thèse
Ahmed TAMTAOUI	PES, Institut National des Postes et Télécommunications, Rabat	Rapporteur/Examinateur
Mohamed El HAZITI	PES, Université Mohammed V-Rabat, École Supérieure de Technologie	Rapporteur/Examinateur
Lahoucine BALLIHI	PH, Université Mohammed V-Rabat, Faculté des Sciences	Rapporteur/Examinateur
Aladine CHETOUANI	PH, Université d'Orléans, Polytech'Orléans, France	Examinateur
Hocine CHERIFI	PES, Université de Bourgogne, Faculté des Sciences Mirande, France	Examinateur

Année Universitaire : 2020-2021

🖂 Faculté des Sciences, avenue Ibn Battouta, BP. 1014 RP, Rabat –Maroc
☎ 00212 (05) 37 77 18 76 📠 00212(05) 37 77 42 61; http://www. fsr.um5.ac.ma

To my parents
To my brothers
To my sister
To my friends

Don't talk like one of them. You're not! Even if you'd like to be.
To them, you're just a freak, like me! They need you right now,
but when they don't, they'll cast you out, like a leper! You see,
their morals, their code, it's a bad joke. Dropped at the first sign
of trouble. They're only as good as the world allows them to be.
I'll show you. When the chips are down, these... these
civilized people, they'll eat each other. See, I'm not a monster.
I'm just ahead of the curve.
The Joker

Remerciements

Cette thèse de doctorat a été menée pour l'obtention du grade de Docteur de l'Université Mohammed V - Rabat. Les travaux présentés dans ce mémoire ont été effectués au Laboratoire de Recherche en Informatique et Télécommunications (LRIT) de la Faculté des Sciences de Rabat (FSR) au Maroc sous la direction du Professeur **Mohammed EL HASSOUNI**.

En premier lieu, je tiens à remercier mon directeur de thèse Professeur **Mohammed EL HASSOUNI**, Professeur d'Enseignement Supérieur à la faculté des lettres et sciences humaines de Rabat, avec qui j'ai travaillé depuis mon projet de fin d'études du Master. Ses conseils précieux ont bien mené à cette thèse. Au-delà de ses qualités scientifiques, il est avant tout un homme généreux, avec qu'il est agréable de discuter, d'échanger des idées, et de plaisanter. J'espère que ces immenses qualités continueront à m'influencer pour longtemps.

Je tiens à remercier Professeur **Mohamed OUADOU**, Professeur d'Enseignement Supérieur à la faculté des sciences de Rabat, d'avoir accepté de présider le jury de ma thèse.

Je tiens aussi à remercier Professeur **Ahmed TAMTAOUI**, Professeur d'Enseignement Supérieur à l'institut national des postes et télécommunications, Rabat, d'avoir accepté de juger la qualité de mon travail en tant que rapporteur et examinateur.

Je voudrais également remercier Professeur **Mohamed EL HAZITI**, Professeur d'Enseignement Supérieur à l'école Supérieure de Technologie de Salé, qui m'a fait l'honneur de rapporter et examiner cette thèse.

Mers remerciements vont également à Professeur **Lahoucine BALIHI**, Professeur Habilité à la faculté des sciences de Rabat, d'avoir accepté de rapporter et examiner ce mémoire de thèse.

Je suis également reconnaissant au Professeur **Hocine CHERIFI**, Professeur à l'Université de Bourgogne, France, pour ses conseils et sa disponibilité. Je le remercie aussi d'avoir accepté de participer en tant qu'examinateur.

Je suis particulièrement redevable au Professeur **Aladine CHETOUANI**, Professeur Habilité à l'Université d'Orléans, France, pour son engagement. Ces conseils et son talent m'inspireront certainement tout au long de ma carrière. Je le remercie aussi d'avoir accepté de participer en tant qu'examinateur.

Je remercie mes amis qui m'ont apporté leur soutien et leurs encouragements tout au long de cette aventure. Dans le désordre, Mohamed(s), Mehdi, Simo, Ayoub, Youssef, Mounir(s), Bethaina, Zakariya et Lamiae.

Je garde le meilleur pour la fin, ma famille qui a supporté toutes les difficultés morales et matérielles pour me soutenir au terme de mes études. J'adresse ma profonde gratitude et mon immense reconnaissance à mes raisons d'être, ma mère, ma mère, ma mère et mon père, qui m'ont éduqué et orienté. Merci de m'avoir encouragé et soutenu dans mes choix. Nul mot et nulles expressions refléteront le grand amour et la profonde gratitude que je porte pour vous. Mes plus sincères remerciements vont aussi à mes frères Jawad Bilal, Bader Pedro, Alae Alwan et ma soeur Salima. Aucun mot ne saurait retranscrire ici sans l'affaiblir le bonheur qu'ils m'ont toujours apporté, ni l'ampleur de ce que je leur dois.

Table des matières

ABSTRACT

As like images and videos, the perceptual quality of 3D meshes can be affected by several external factors (watermarking, compression, simplification, etc.). To estimate the impact of these treatments, several quality measures have been proposed in the literature. There are currently three main families of metrics : full reference (FR) which assumes that the reference mesh is available, reduced reference (RR) which only uses extracted features from the reference mesh, and no reference measures (NR), also called blind, which do not have access to any information oof reference. A certain full or reduced reference methods have been proposed to estimate the perceived visual quality of 3D meshes. However, in most practical situations, access to information about the reference and the type of distortion is limited. For these reasons, the development of a no-reference method is a crucial issue. In this context, we are interested in this thesis in no-reference mesh visual quality assessment. The first contribution concerns the development of a method based on feature learning approaches to predict the quality scores. This method uses features extracted from the distorted mesh and feature learning methods for quality estimation. As for the second contribution, deep learning is used to assess the visual quality. We begin by a simple CNN network from scratch and then we take advantage of deeper networks with a combination based on the Compact Multi-linear Pooling (CMP). In the third contribution, the 3D visual saliency is used to prepare the learning data. CNN architectures are fed by small patches carefully selected according to their level of saliency. In the last contribution, we introduce a "model-based" method using convolutional graph networks (GCN) to directly process the 3D model itself without using 2D patches rendered from the 3D model (image-based).

Keywords : Triangular mesh. Blind mesh quality assessment. Human visual system. Objective evaluation. Subjective evaluation. Perceptual quality. Deep learning. Visual saliency. Convolutional neural network.

Résumé

De même que les images et les vidéos, la qualité perceptuelle des maillages 3D peut être affectée par plusieurs facteurs externes (tatouage, compression, simplification, etc.). Afin d'estimer l'impact de ces traitements, plusieurs mesures de qualité ont été proposées dans la littérature. Ils existent actuellement trois grandes familles de métriques : référence complète (FR) qui suppose que le maillage de référence est disponible, référence réduite (RR) qui exploite uniquement des caractéristiques du maillage de référence et les mesures sans référence (NR), appelées également aveugle, qui n'ont accès à aucune information du maillage de référence. Un certain nombre de méthodes à référence complètes ou réduites ont été proposées afin d'estimer la qualité visuelle perçue des maillages 3D. Cependant, dans la plupart des situations pratiques, l'accès aux informations relatives à la référence et au type de distorsion est limité. Pour ces raisons, le développement d'une méthode de qualité visuelle sans référence est une problématique cruciale. Dans ce cadre, nous nous intéressons dans cette thèse à l'évaluation sans référence de la qualité perceptuelle des maillages 3D. La première contribution porte sur la proposition d'une méthode basée sur une approche d'apprentissage afin de prédire les scores de qualité. Cette méthode utilise des caractéristiques extraites à partir du maillage déformé et des méthode d'apprentissage pour l'estimation de la qualité visuelle. Quant à la seconde contribution, l'apprentissage profond est exploité pour estimer la qualité. Nous commençons par un réseau CNN simple et ensuite nous utlisons des réseaux plus profond avec une combinaison basée sur le Compact Multi-linear Pooling (CMP). Dans la troisième contribution la saillance visuelle 3D est exploitée pour préparer les données d'apprentissage. Les architectures CNN sont alimentées par des petits patchs soigneusement sélectionnées en fonction de leur niveau de saillance. Dans la dernière contribution, nous introduisons une méthode "model-based" utilisant des réseaux convolutionnels de graphes (GCN) pour traiter directement le modèle 3D lui-même sans utiliser des patchs rendus (image-based).

Mots clés : Maillage triangulaire. Évaluation sans référence de la qualité des maillage 3D. Système visuel humain. Évaluation objective. Évaluation subjective. Qualité perceptuelle. L'apprentissage profond. Saillance visuelle. Réseau de neurones convolutifs.

Résumé détaillé

La qualité perçue de maillages 3D est le résultat de différentes opérations liées à la transmission et le traitement géométrique (tatouage, simplification, compression, ...). La qualité perceptuelle d'un maillage 3D est subjectivement défini comme la moyenne des évaluations effectuées par des sujets humains (MOS : Mean Opinion Score). Cependant, l'évaluation subjective est coûteuse, laborieuse et prend un certain temps. Les méthodes objectives d'évaluation sont la solution la plus adéquate pour évaluer automatiquement la qualité visuelle. Le problème de l'évaluation de la qualité visuelle des maillage 3D a connu des progrès considérables au cours des dernières années. Les premiers travaux ont utilisé des similarités simples entre le maillage de référence et sa version déformée telle que l'erreur quadratique moyenne (RMS) et la distance de Hausdorff (HD). Ce type de méthodes a généralement échoué car il calcule une distance géométrique négligeant les opérations principales du système visuel humain (SVH) . Afin d'intégrer l'information perceptuelle, plusieurs méthodes utilisent différents principes pour une meilleure estimation de la qualité visuelle perçue. Dans , une métrique perceptuelle basée sur l'analyse de courbure appelée mesure de distorsion structurelle de maillage (MSDM) a été proposée. Afin d'évaluer la qualité des maillages tatoués, Corsini et al. ont développé une métrique perceptuelle utilisant la variation de rugosité . Une autre mesure perceptuelle appelée FMPD (Fast Mesh Perceptual Distance) a été proposée dans . Cette métrique est basée sur une mesure de rugosité locale dérivée de la courbure gaussienne. Ces méthodes, avec référence, atteignent une corrélation très élevée avec la perception humaine. Cependant, leur principal inconvénient est l'indisponibilité du maillage de référence dans les applications réelles.

Cadre général et objectif

Ce mémoire intitulé *Évaluation sans référence de la qualité des maillages 3D : Exploitation de l'apprentissage profond et la saillance visuelle* est présenté en vue d'obtenir le titre de docteur de l'Université Mohammed V-Rabat, dans la spécialité ”Informatique et Télécommunications. Ce travail est dans le domaine de l'évaluation de la qualité visuelle des maillages 3D.

L'objectif principal est de mettre en oeuvre des méthodes automatisées assurant l'opération de prédiction des scores de qualité en exploitant des techniques d'apprentissage y compris l'apprentissage profond. Le score prédit (évaluation objective) par une méthode automatique doit assurer une bonne corrélation avec l'évaluation subjective fournie par les observations humaines (évaluation subjective).

Dans ce cadre nous avons proposé quatre nouvelles approches. La première concerne le développement d'une méthode basée sur des approches d'apprentissage et des caractéristiques extraites du maillage déformé. Dans la deuxième, l'apprentissage profond, est utilisé pour évaluer la qualité visuelle. Nous commençons par un réseau CNN de base, puis nous profitons des réseaux plus profonds avec la combinaison multilinéaire compacte (CMP). Dans la troisième contribution, la saillance visuelle 3D est utilisée pour préparer les données d'apprentissage. Les architectures CNN sont alimentées par des patchs soigneusement sélectionnés en fonction de leur niveau de saillance. Dans la dernière contribution, nous introduisons une méthode *model-based* en utilisant un réseau de graphes convolutifs (GCN) pour traiter directement le maillage 3D sans utiliser de patchs 2D (image-based).

Organisation du manuscrit

Le mémoire est constitué d'une introduction, suivie de cinq chapitres principaux et d'une conclusion et des perspectives.

Introduction générale : donne un aperçu sur le sujet traité dans la thèse. Elle relate le contexte du travail en mettant en avant tout l'intérêt de l'évaluation de la qualité visuelle. Ensuite, nous présentons la motivation et l'objectif de la thèse ainsi que le plan et un résumé de l'ensemble des contributions.

Chapitre 1 : État de l'art sur la qualité perceptuelle des maillages 3D

Dans ce chapitre, nous présentons les propriétés les plus importantes du système visuel humain et sa relation avec l'évaluation de la qualité. Ensuite, nous décrivons l'évaluation subjective et les différents paramètres utilisés dans une expérience subjective. Ensuite, des méthodes objectives dans l'état de l'art sont analysées, y compris les méthodes à référence complètes, à référence réduites et sans référence. Des bases de données subjectives spécialement conçues pour l'évaluation de la qualité visuelle sont également abordées. Enfin, nous décrivons le protocole de validation et la configuration expérimentale qui sert à l'évaluation des performances d'une méthode donnée.

Chapitre 2 : évaluation de la qualité visuelle basée sur des caractéristiques perceptuelles et des méthodes d'apprentissage

Dans ce chapitre, nous introduisons deux caractéristiques perceptuelles utilisées pour décrire le maillage 3D. Nous adoptons la courbure moyenne qui représente l'aspect visuel, et l'angle dièdre qui représente l'aspect structurel du maillage 3D. Deux méthodes d'apprentissage notamment Support vector regression (SVR) et General regression neural network (GRNN) sont utilisées pour estimer le score de qualité en se basant sur les caractéristiques perceptuelles extraites à partir du maillage déformé.

Chapitre 3 : réseaux de neurones convolutifs pour l'évaluation de la qualité visuelle

Dans ce chapitre, nous présentons tout d'abord l'apprentissage profond, en particulier le réseau de neurones convolutifs et ses applications. Ensuite, nous décrivons comment on a préparé les données d'entrée utilisées dans notre approche. Une description détaillée de la première méthode basée sur un CNN utilisant un réseau basique est présentée. Ensuite, nous présentons une méthode développée qui utilise des réseaux pré-entrainé et un regroupement multilinéaire compact Compact multilinear pooling (CMP) pour l'estimation de la qualité.

Chapitre 4 : Effet de la saillance visuelle sur l'estimation de la qualité perceptuelle

Dans ce chapitre, nous étudions l'utilisation des architectures CNN et de la saillance visuelle 3D pour estimer la qualité visuelle perçue des maillages déformés. Il s'agit d'une extension des tests réalisés dans le chapitre 3. La valeur ajoutée est d'étudier l'effet de la saillance visuelle dans notre démarche d'estimation de la qualité à l'aide des CNN. Pour ce faire, l'architecture CNN est alimentée par de petits patchs sélectionnés en fonction de leur niveau de saillance.

Chapitre 5 : Réseau convolutif sous forme de graph (GCN) pour l'estimation de la qualité visuelle

Dans ce chapitre, nous présentons une nouvelle approche pour estimer la qualité perçue des maillages 3D. Cette méthode est sans référence et elle est basée sur un réseau convolutif de graph avec trois modules (convolution, pooling et classification) pour s'appuyer sur le problème de la classification des noeuds et estimer le score de qualité perçu. Cette approche profite de l'opération de convolution en utilisant directement le maillage 3D lui-même sans rendre les vues 2D. Le graphe se compose de deux couches de convolution, une couche de pooling regroupement et une couche de classification. Le réseau est alimenté par un graphe représenté par une matrice d'adjacence et une matrice de caractéristiques contenant quatre caractéristiques géométriques (courbure, angles, Laplacien de courbure Gaussienne et la saillance). Nous avons testé plusieurs caractéristiques avant d'arriver à configuration la plus adéquate.

Résultats expérimentaux

Pour tester la performance d'une méthode de qualité visuelle, une base de données de maillages dégradés notés par des observateurs humains est nécessaire. Notre méthode d'évaluation de qualité a été testée et validée à l'aide de deux bases de données accessibles au public spécialement conçus pour l'évaluation des méthodes de qualité :

- LIRIS Masking : contient 4 modèles de référence et 24 modèles déformés obtenus par l'addition locale de bruit avec différents niveaux.
- General-purpose : contient 4 modèles de référence et 84 modèles déformés obtenus par l'addition locale du bruit et le lissage avec différents niveaux.
- UWB compression : contient 5 modèles de référence et 63 modèles déformés (68

modèles au total), pour chaque modèle de référence, douze ou treize versions déformées sont crées.

— IEETA simplification : contient cinq modèles de référence et 30 versions simplifiées (six versions déformées pour chaque référence). Les modèles simplifiés ont été obtenus à l'aide de trois algorithmes de simplification avec deux rapports de réduction de sommets différents.

Pour évaluer la performance des méthodes proposées, deux coefficients de corrélation sont couramment utilisés, à savoir le coefficient de corrélation linéaire de Pearson (rp) (précision de prédiction) et le coefficient de corrélation de Spearman (rs) (monotonie de prédiction).

Le tableau ci-dessous présente les coefficients de corrélation (rs) et (rp) des méthodes comparées obtenus sur la base de données General-purpose. Les corrélations sur l'ensemble du corpus sont calculées entre les scores objectifs de tous les objets dans le corpus et leurs MOS correspondants.

Tableau 1 – Les coefficients de corrélation r_s (%) and r_p (%) des méthodes comparées obtenus sur la base de données General-purpose.

Type	Metric	Armadillo		Dyno		Venus		Rocker		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	69.5	30.2	30.9	22.6	1.6	0.8	18.1	5.5	13.8	1.3
	RMS Cignoni *et al.* (1998)	62.7	32.3	0.3	0.0	90.1	77.3	7.3	3.0	26.8	7.9
	MSDM2 Lavoué (2011)	81.6	85.3	85.9.4	85.7	89.3	87.5	89.6	87.2	80.4	81.4
	TPDM Torkhani *et al.* (2014)	84.5	78.8	92.2	89.0	90.6	91.0	92.2	91.4	**89.6**	**89.2**
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	65.8	35.7	62.7	35.7	71.6	46.6	87.5	53.2	69.3	38.4
	3DWPM2 Corsini *et al.* (2007)	74.1	43.1	52.4	19.9	34.8	16.4	37.8	29.9	49.0	24.6
	FMPD Wang *et al.* (2012)	75.4	83.2	89.6	88.9	87.5	83.9	88.8	84.7	**81.9**	**83.5**
No-Reference	NR-SVR Abouelaziz *et al.* (2016b)	76.8	91.5	78.6	84.1	85.7	88.6	86.2	86.6	81.5	87.8
	NR-GRNN Abouelaziz *et al.* (2016a)	87.1	**97.3**	91.2	**94.1**	86.3	85.0	78.6	74.8	86.2	88.7
	NR-CNN1 Abouelaziz *et al.* (2017)	87.2	84.3	86.4	86.2	92.2	85.6	91.3	85.2	83.6	82.7
	NR-CNN2 Abouelaziz *et al.* (2018)	93.4	95.6	86.2	84.3	**94.1**	90.3	80.4	82.2	81.8	82.5
	BMQI Nouri *et al.* (2017)	20.1	-	83.5	-	88.9	-	92.7	-	78.1	-
	CNN-BMQA	90.6	92.4	88.3	86.3	93.1	92.6	88.9	89.0	**90.0**	**92.0**
	CNNs-CMP	93.4	92.9	91.6	90.9	88.9	89.4	92.6	93.9	**92.6**	**91.3**
	SCNNs-CMP	**95.8**	95.6	**93.6**	92.9	93.4	**91.3**	**94.5**	**95.2**	94.4	94.8
	MVQ-GCN	91.8	92.5	87.7	84.5	93.7	91.9	89.6	88.4	**89.3**	**88.6**

les résultats ont montré que les méthodes basées sur les distances géométriques ne reflètent pas la qualité perçue et ne correspondent pas bien avec la perception humaine. D'autre part, les méthodes perceptuelles et les méthodes proposées atteignent des corrélations élevées et montrent une bonne performance dans l'évaluation de la qualité perçue. Les méthodes proposées fournissent des bonnes corrélations sur les quatre bases des données. Il convient de noter que les méthodes proposées sont sans référence et ne nécessite pas le maillage de référence. Contrairement aux méthodes de référence complètes et réduites concurrentes, nos méthodes peuvent être utiles dans des situations pratiques.

List of abreviations

MVQ	*Mesh Visual Quality*
FR	*Full Reference*
RR	*Reduced Reference*
NR	*No Reference*
CAD	*Computer Aided Design*
HVS	*Human Visual System*
RMS	*Root Mean Square*
SVR	*Support Vector Regression*
GRNN	*General Regression Neural Network*
CNN	*Convolutional Neural Network*
CSF	*Contrast Sensitivity Function*
NSS	*Natural Scene Statistics*
SSIM	*Structural Similarity Metric*
MSDM	*Mesh Structural Distortion Measure*
MOS	*Mean Opinion Score*
RBF	*Radial Basis Function*
LOOCV	*leave-One-Out Cross Validation*
CMP	*Compact Multi-linear Pooling*
ANNs	*Artificial Neural Networks*
ILSVRC	*Image Large Scale Visual Recognition Challenge*
VGG	*Visual Geometry Group*
ResNet	*Residual Neural Network*
CRF	*Conditional Random Field*
3DWPM	*3D Watermarking Perception Metric*
FMPD	*Fast Mesh Perceptual Distance*
TPDM	*Tensor-based Perceptual Distance Metric*

Table des figures

Liste des tableaux

Introduction

Mesh representation

Technological advances in computer graphics, telecommunication, and computer-aided design over the two past decades have contributed to the development of three-dimensional (3D) data. Thanks to the rapid development of hardware and software for both professionals (3D modeling tools) and users (graphic cards, smart-phones capable of viewing 3D models), 3D data is widely used nowadays. As shown in Fig. 1, several application fields are directly concerned by this type of data such as computer-aided-design (CAD), architecture, cultural heritage (3D digitizing of the ancient statues), digital recreation (video games, 3D movies), scientific visualization and so forth.

Figure 1 – Several 3D models from different fields of application. from left to right : gaming character (digital leisure) a scanned statue (cultural heritage), a mechanical part (CAD)

3D data (describing a human, animal, or an object) can be represented in different ways ; however, in most applications, it is represented by polygonal meshes that model the surface of objects by a set of vertices and faces. This representation, particularly triangular meshes, is widely used rather than other surface models such as implicit surfaces or parametric surfaces. It presents a numeration of the object or a 3D scene.

Depending on the geometric shape of the faces, we can distinguish several types of 3D meshes Botsch *et al.* (2010). As illustrated in Fig. 2 a triangular polygonal mesh M is defined by its geometry (vertices) and its connectivity which describes the relations between the vertices of the mesh. The elements composing a mesh M are :

— A set of vertices $S = \{s_1, ..., s_n\}$
— A set of triangular faces $F = \{f_1, ..., f_m\}$, $f_m \in S^3$
— A set of edges $A = \{a_1, ..., a_l\}$, $a_i \in S^2$

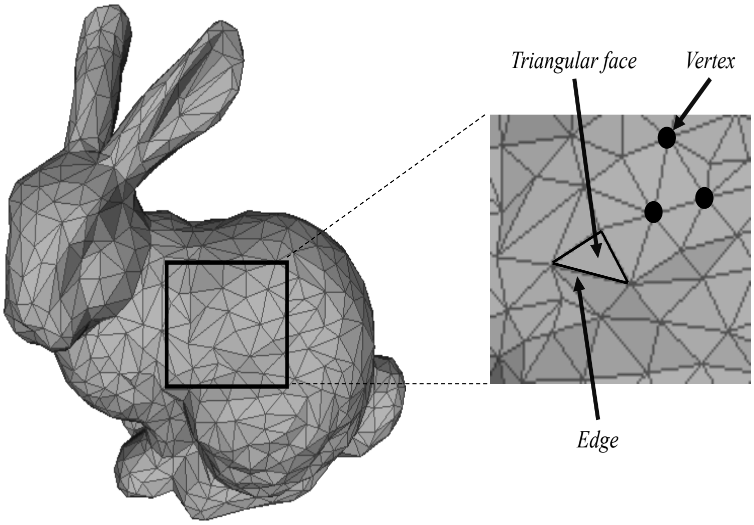


FIGURE 2 – Elements of a triangular 3D mesh

3D objects can be generated by an acquisition method (reality) : scanner 3D, multiscopic vision, MRI, ultrasound, etc. Or by creating a graphical object using a geometric parametrization tool (synthesis). Fig. 3 shows the life cycle of a 3D mesh from the acquisition to the visualization.

Geometric processing

In practical situations, 3D meshes are usually subject to different lossy geometry processing operations such as watermarking Wang *et al.* (2008); Wang et Hu (2009), simplification Garland et Heckbert (1997); Luebke (2001), and compression Alliez et Gotsman (2005); Lee *et al.* (2011).

— Compression : the goal is to reduce the size of the large 3D data in order to facilitate the storage and reduce the transmission time. It is also performed to adapt the complexity of the 3D models to the constraints of the visualization, in particular in the context of multi-resolution methods, in order to be able to offer an experience of interaction and real-time visualization.
— watermarking : 3D objects can be easily copied and redistributed by unauthorized users. As a result, the need to protect these contents has increasingly become crucial. Digital watermarking, which consists of inserting a watermark in the host

data to protect copyright, is considered as an efficient solution to overcome the above-mentioned issue. the objective here is to protect the intellectual property of 3D content, it can be hacked during its transmission on the network or its display on a client terminal. Digital watermarking consists of inserting an invisible mark in the data, most of the time by intelligently modifying the geometry, and thus makes it possible to carry out the proof of ownership. In addition to copyright protection, this technique also makes it possible to track copies or even verify the integrity of data.

— Simplification : Complex objects simulate the reality very well but have as disadvantages difficulty in handling, rendering, or transmitting over the internet. Moreover, the storage memory is larger for complex objects. For these reasons, mesh simplification is desirable. The goal of the mesh simplification algorithms is to reduce the complexity of a mesh while preserving a high fidelity of the original during simplification.

— Remeshing : some complex treatments which may be done on 3D meshes need particular connectivity (regular connectivity with a defined valence or connectivity of . subdivision, etc.). A pre-treatment is so necessary : the original object will be remeshed to have the needed connectivity. This operation also modifies the representation of the mesh.

In addition, the 3D meshes may subject to random noise during the transmission process, which affects the mesh and add undesirable information.

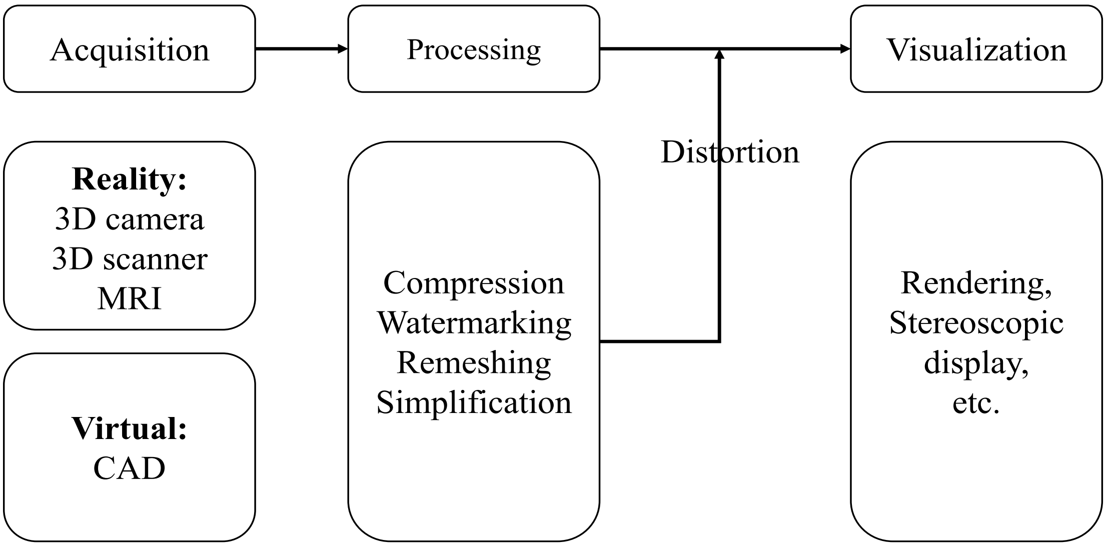


Figure 3 – From the acquisition to the display of 3D meshes.

Quality assessment

The geometric processes inevitably introduce variable modifications on the 3D shape of the object that may alter the visual quality of the model. The modification is called distortion In all possible scenarios, it is crucial to evaluate how much the original 3D model has been affected by a specific operation. Mesh visual quality (MVQ) assessment is the field of study that tries to identify how much the original 3D model has been distorted. The study of perceptual quality is an important task since most visual data is intended for an ultimate human observer. In fact, visual distortions can be measured by human observers (subjective visual quality assessment), but this evaluation is too expensive, laborious, and time-consuming compared to the objective visual quality assessment that seems to be a good solution to overcome these problems. It is thus indispensable to implement automatic methods to objectively assess the visual quality and ensure a good correlation with the subjective assessment provided by human observations. For each observer, the quality can have a different definition according to personal criteria. Objective methods can be classified according to the availability of the reference : full-reference (FR) when the reference model is fully available, reduced-reference (RR) when only a part of the reference is available and no-reference (NR) when no information of the reference is available. Objective methods can also be categorized into two groups : image-based and model-based (or geometry-based). In the first category, the perceptual mechanisms are applied to the images generated from the 3D data while in the second group the perceptual metrics work directly on the 3D model itself making the evaluation view-independent. Several studies on the measurement of objective perceptual quality have been carried out. The main objective of these studies is to analyze the behavior of the human visual system and to quantify the quality as perceived by the user of this multimedia data. Several research teams are interested in the problem of objective evaluation of the perceptual quality of multimedia content (sound, image, 2D/3D video, and three-dimensional meshes). The objective of perceptual quality metrics is to judge with certainty the quality of a mesh with regard to human perception. To do so, there are different approaches : Top-down and Bottom-up approaches. The first category of metrics considers the HVS as a black box and tries to imitate the behavior of this system from the input/output point of view. The second category is based on the simulation and imitation of each component of the HVS. To compare two 2D images or two 3D meshes, it is common to use measures called signal to noise ratio or geometric distances. These geometric measurements do not integrate the properties of the human visual system. We present in Fig. 4 an example of a reference mesh and two distorted versions, D_2 is with higher distortions level (noise addition). It is remarkable that the version D_1 has a better perceptual quality. However, the objective metric Root mean square error (RMS) provides exactly the same geometric distance value for the two versions of each mesh with respect to the reference mesh. This example shows the importance of deriving perceptual metrics capable of replacing these distances which are not capable to judge the perceived quality of the degraded meshes with accuracy.

Contributions

The blind quality assessment has been successfully adopted for image and video processing, many methods in this field of study have shown notable results in terms of the correlation with human judgment. While the studies in mesh quality assessment tend

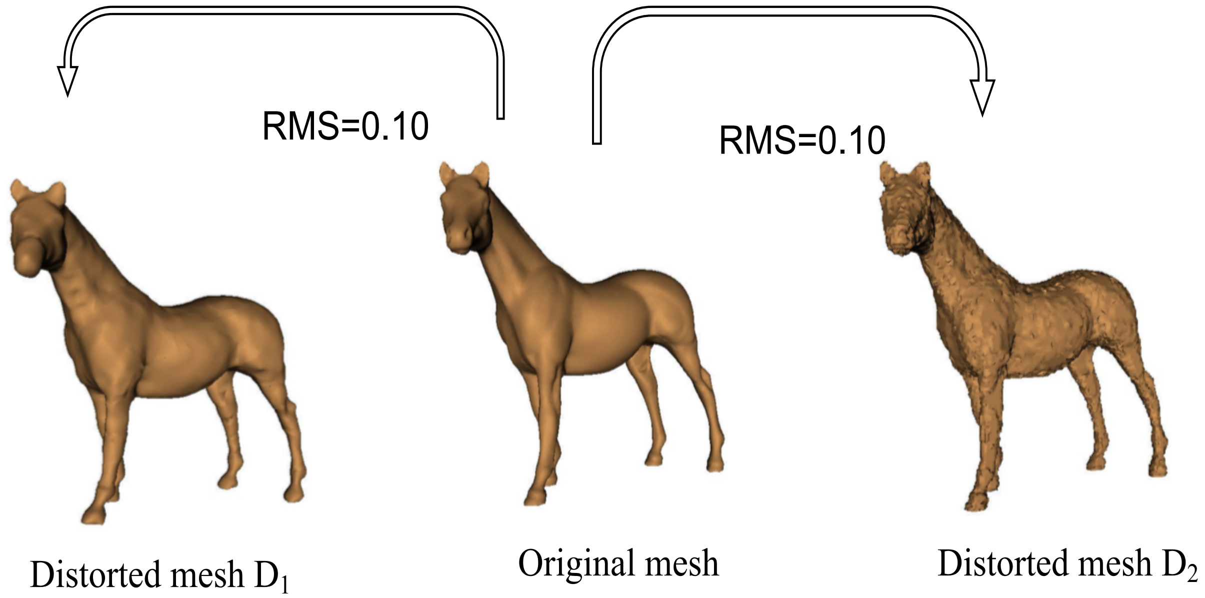


FIGURE 4 – Elements of a triangular 3D mesh

more to adopt full-reference and reduced-reference approaches. However, it is required to address more importance to blind MVQ assessment. In this context, we focus on blind MVQ assessment in order to handle the practical situations where the reference mesh is not available. Our contributions can be summarized as follows :

- The first contribution relies on a feature learning-based approach to predict the objective quality scores. For this, we propose the mesh dihedral angles statistics as a handcrafted feature and the support vector regression (SVR) as a learning tool based quality predictor. The proposed method takes into account the main functions of the human visual system (HVS) by introducing the visual masking. We used also the mean curvature which considered as a perceptual relevant feature representing the visual aspect of a 3D mesh. The general regression neural network (GRNN) is adopted to learn the mean curvature and predict the perceived visual quality
- In the second contribution, we propose a convolutional neural network (CNN) architectures to estimate the perceived visual quality of 3D meshes. We decided to use a deep network instead of a classical regressor used in the first method. First, we use a network from scratch, then we rely on pre-trained deep convolutional neural networks VGG, AlexNet, and ResNet. To do so, the CNN architectures is fed by small patches, we render 2D projections from the 3D mesh. Then the obtained views are split into 2D small patches. The CNN from scratch with convolution and max-pooling layers is used for the feature learning and the quality score estimation. For the pre-trained networks, Each network is fine-tuned and produces a feature vector. The compact multi-linear pooling is then used to fuse the extracted feature vectors into a global feature representation. The main objective is to use deeper architectures and an interesting combination strategy to take advantage of many networks in parallel.

— In the third contribution, we extend the methods based on CNNs by including 3D visual saliency. The HVS is supposed to have more sensitivity to distortions in salient regions. Thus, we use mesh saliency to indicate the most relevant regions of the 3D mesh. A patch selection technic based on mesh visual saliency is used to select only the relevant patches to feed the CNN architectures i.e. the network from scratch and three pre-trained networks : VGG, AlexNet, and ResNet. Several tests will be conducted to prove the importance of the patch selection strategy.

— In the fourth contribution, we use a model-based method using Graph Convolutional Networks (GCN) to work directly on the 3D model itself. GCNs has been mainly applied for node classification tasks in which the convolution representation vector for a node function as the only features to classify that node. For the mesh quality assessment task, the 3D model is represented by a graph, and the graph-based convolution vector of nodes is used to predict the perceived visual quality. The main objective of this contribution is to use directly the 3D mesh without rendering 2D views.

Thesis plan

This dissertation is organized as follows :

— Chapter 1 presents the state of the art on 3D mesh visual quality assessment. This chapter states the existing objective metrics for estimating the quality of 3D meshes. It introduces the perceptual quality and the human visual system. In addition, a detailed description is given of the subjectively-rated mesh visual quality databases and the validation protocol to test the performance of MVQ methods.

— Chapter 2 presents a perceptual quality estimation based on handcrafted features and machine learning techniques.

— Chapter 3 introduces a deep learning approach relying on a convolutional neural network architecture from scratch and pre-trained networks with a combination technique to fuse the feature vectors.

— Chapter 4 presents the use of 3D visual in mesh quality assessment. we include a saliency patch selection based on the level of saliency and study the effect of this perceptual parameter.

— Chapter 5 presents a graph convolutional network to predict the visual quality, in this chapter we introduce the notion of graphs to use directly the 3D shape (model-based).

— Finally, in the general conclusion we present a summary of our contributions described in this thesis and we present the perspectives that may be the subject of future research.

Chapitre

1

Perceptual quality of 3D meshes : State of the art

Sommaire

1.1 Introduction

The human visual system is a complex sensory system. Human vision is based on two mechanisms : low-level mechanisms that are the biophysical structure of the sensory system, and high-level mechanisms that relate to the human cognitive system. It can be considered an optical system. When light from an object falls on the eye, the pupil of the eye acts as an aperture, and an image is created on the retina, and the viewer sees the object. The perceived size of an object depends on the angle it creates on the retina. The retina can resolve detail better when the retinal image is larger and spread over more of its photoreceptors. The pupil controls the amount of light entering the eye. For typical illumination, the pupil is about 2 mm in diameter. For low illumination, the size increases to allow more light, whereas for high illumination the size decreases to limit the amount of light.

Recently, experts and researches in computer graphics tried to develop effective objective metrics, able to predict the perceived quality of images, videos, and 3D shapes. To design such a performant system, it is mandatory to study the perceptual structures of the visual content. When it comes to 3D information (represented in the form of triangular meshes)

this task becomes more complex.
Human perception plays an important role to determine the quality of distorted 3D meshes. The subjective evaluation is established by observers through psychometric experiments. This evaluation permits to develop databases with labeled objects to have ground-truth measurements of the perceptual quality of meshes. On the other hand, the objective evaluation tries to maximize the correlations between the predicted and the subjective scores.
The objective of existing metrics is to maximize the correlation of prediction results with subjective scores. In this context, several methods have been proposed that try to mimic an ideal human observer and accurately predict the subjective assessment scores. Some of them use simple distances between the reference and the distorted model, while others tend to develop perceptually driven quality methods in order to take into account the characteristics of the HVS for better estimation.
In this chapter, we first present the most important properties of the human visual system and its relationship to the quality assessment. After that, we describe the subjective evaluation and the different parameters used in a subjective experiment. Next, the state of the art objective methods is analyzed including full, reduced, and no reference methods. Subjective databases specially designed for MVQ assessment evaluation are also discussed. Finally, we describe the validation protocol and the experimental setup that serves in the performance evaluation of a given method.

1.2 The human visual system and the perceptual quality

Understanding human perception and cognition, and modeling the HVS behavior is an essential step for developing quality assessment approaches. This permits us to take advantage of the end-user perception to hide or highlight specific details and thus evaluate the perceived quality of an image, video, or 3D meshes. The HVS perceives a stimulus depending on its color intensity, orientation, and also on its spatial distribution. This important phenomenon caused by the visual cortex allows one to avoid capturing imperceptible information. In this section, we describe the knowledge acquired in the human visual system. We also provide an overview of the main characteristics of human perception that have been widely exploited in recent years. It allows one to have a better understanding of the major phenomena of the human visual system (HVS), such as sensitivity to contrast, visual masking, and so on.

1.2.1 Description of the visual system

The visual information is received by the eye, the brain interprets information after various preliminary treatments carried out by the retina. In fact, higher-level treatments are realized in the brain at the level of the visual areas found in the rear part of the two cerebral hemispheres. According to physiological studies Bullier (1998), an important place has been transferred for the cortical zones dedicated to the processing of visual information. Besides, the links that connect with other areas of the brain make the visual system very complex. Fig. 1.1 describes the main treatments : retinal treatments, the transmission of information via the optic nerve, and treatments performed by the first area of the primary visual cortex.

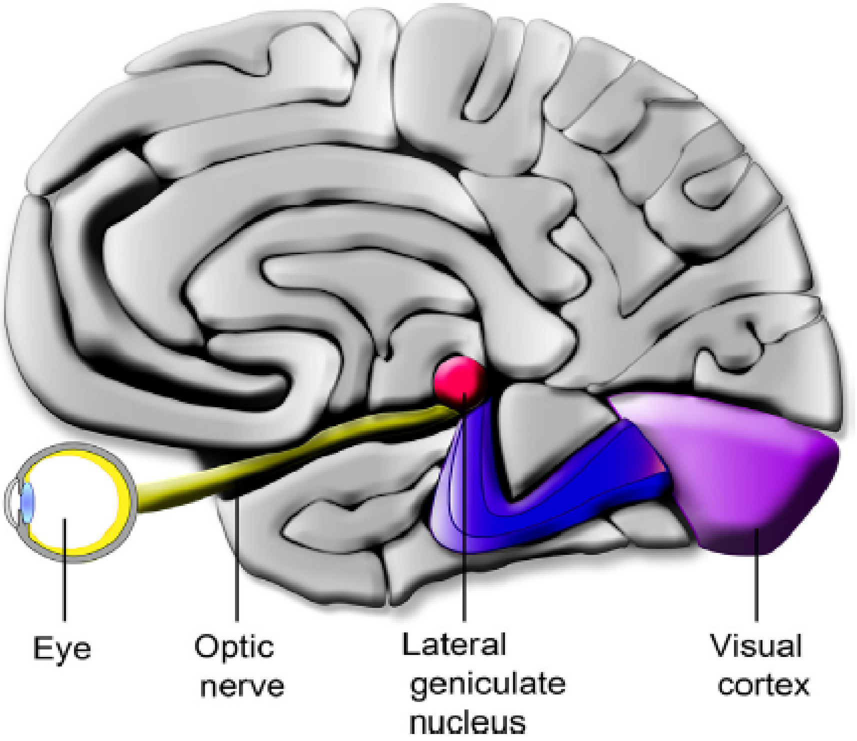


FIGURE 1.1 – Global diagram of the human visual system.

The retina

the retina represents a neurosensory layer in the eye, it transforms the light received into chemical or electrical signals, which are then transmitted to the visual cortex by the optic nerve Denis (2014). The retina is made up of three layers of cells. Fig. 1.2 illustrates Structure of the retina layers.

— **The layer of ganglion cells** constitutes the optic nerve : a million neurons come together to form this nerve, these neurons connect the lower layer to the brain.

— **The inner granular layer** has three types of cells. information is transmitted by bipolar cells between the two main poles of the retina from the layer of photoreceptors made up of rods and cones to the optic nerve. First, the nervous message is adjusted according to the horizontal cells by modulating the visual information taking into account the surrounding colors. Then, the adjustment of the nervous message continues by modulating the visual information taking into account the brightness, contours, and color mixing of the objects.

— **The photoreceptor layer** ensures the transmission of the light signal into an electrical signal. This involves 136.5 million cells, cones, and rods.

The visual cortex

The behavior of the visual cortex remains vague. However, studies on primates have led to some major advances without these results being systematically transferable to the

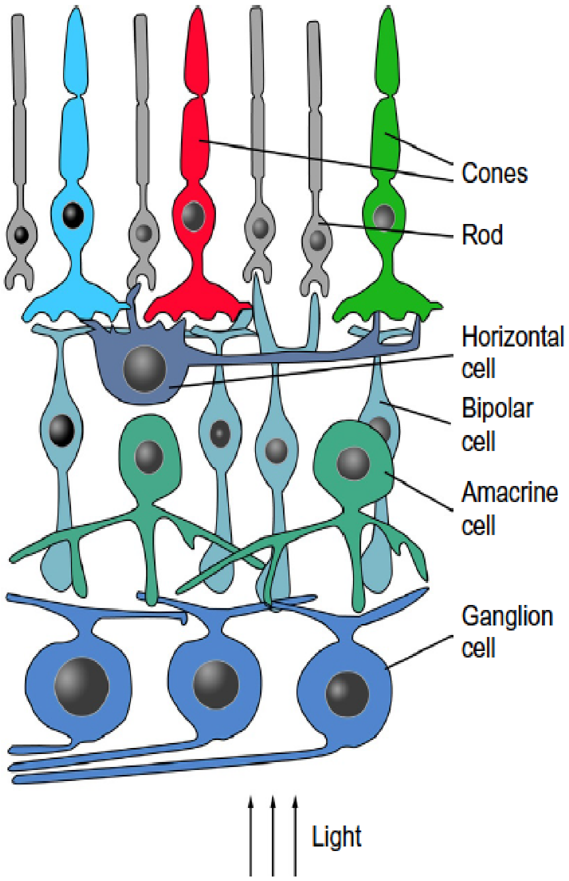


FIGURE 1.2 – Structure of the retina layers.

function of the human visual system. This work highlighted several challenges to be met and made it possible to define these three characteristics : retinotopy, the organization in layers and organization in columns :

- Retinotoy indicates that the foveal surface of the retina, which represents a small proportion of the visual field, projects onto almost half of the cortical surface of area V1. The remaining half of the cortical area is occupied by the other part of the visual field of the retina, which shows that the visual capacities are degraded there.
- the second characteristic of the visual cortex is the organization in layers. Each of the six layers is more or less specialized in aspects of information processing with a notation from I to VI. Layer IV which is subdivided into sublayers receives information from the Lateral geniculate nucleus.
- Hubel et al. Hubel et Wiesel (1962) describe the organization in columns. They have shown that the vertical processing of the association is combined with the layer organization described above. In addition, each column of the visual cortex processes a characteristic of the information : luminance, contrast, color, movement, shape, etc. In addition, the influence of spatial orientation has been particularly studied. This study was carried out by presenting to monkeys stimuli from

slits or inclined edges of the light. It leads to the conclusion that a deviation of 10 degrees causes significant variations.

1.2.2 Characteristics of human perception

Contrast sensitivity function

The contrast sensitivity function (CSF) is used in the construction of many metrics and systems in the imaging field. Assuming that an observer can be in a range of viewing distances, the CSF can be considered as the envelope of functions each describing the behavior for a given range of spatial frequencies.

Psychometric saturation effect

The psychometric saturation effect means that when asked to assess the intensity of a stimulus, humans tend to provide a constant response at extreme quantities that are beyond or below a certain threshold. Typically, human observers do not distinguish between very small (or large) stimuli of slightly different intensities, and observers will assign the same subjective score for slightly different but very bad (or good) qualities.

Visual masking effect

The visual masking effect is one of the most important characteristics of the HVS. It takes place when the visibility of a signal is reduced or *masked* by the presence of another signal, called a mask. The contrast or lightness is reduced by different levels that might arrive at invisibility. The amplitude of this effect is measured by the variation of the visibility of the masked signal before and after the presence of the masking signal. Three types of visual masking according to the timing arrangements : simultaneous masking, when the target and mask appear together. Forward/backward masking when the mask precedes/follows the target.

In the context of 3D modeling, the concept of visual masking can be explained by the fact that the human perception cannot notice a small distortion located in a rough area, whereas human observers can detect a distortion easily in smooth areas. Fig. 2.2 illustrates a concrete example of the visual masking effect. Fig. 2.2(a) and Fig. 2.2(b) represent respectively the original LionVase model and its corresponding roughness map. As highlighted, two remarkable regions could be distinguished in the 3D model : rough regions, and smooth regions. Fig. 2.2(c) represents a distorted version of the LionVase model obtained by introducing high-level noise in rough regions, and Fig. 2.2(d) shows another distorted version obtained by adding medium-level noise in smooth areas. Subjectively, it is obvious that the perceived quality in Fig. 2.2(b) is better than the perceived quality in Fig. 2.2(c). Although the objective level of distortion in Fig. 2.2(b) (high-level noise) is higher than that in Fig. 2.2(c) (medium-level noise), since the introduced noise hardly affect smooth regions, while it is less annoying in rough regions.

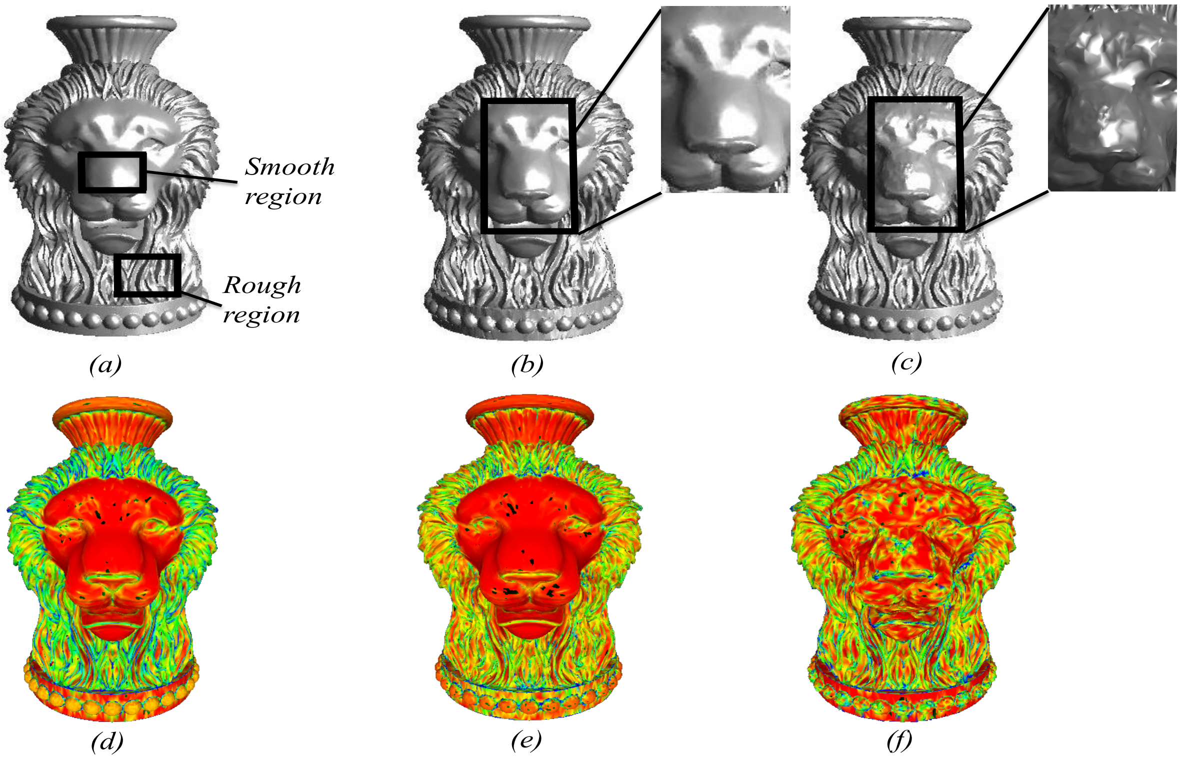


FIGURE 1.3 – An example of the visual masking effect. (a) Original LionVase model and illustration of the different regions types ; (b) high-level noise applied on rough regions ; (c) medium-level noise applied on smooth regions ; (d) roughness map of the original model : rough regions shown with warmer colors ; (e) roughness map of model in "b" ; (f) roughness map of the model in "c".

Visual saliency

Visual saliency is an important aspect of human perception. It has been widely exploited recently in computer vision applications. This property is crucial to explore the surrounding visual world and it is considered one of the most important elements in human perception. To understand and study this aspect, researchers tend to use eye-tracking. Yarbus Yarbus (2013) performed a study in 1960s. It is one of the most popular studies that try to analyze the eye movement. To do so, Ilya Repin's painting was shown to many observers and different viewing tasks were assigned. Fig. 1.4 illustrates The visual paths of these observers. It is noted that the observations of images translate into a sequence of saccades and fixations on key/interest points of the observed object.

1.3 Subjective evaluation

Subjective evaluation is performed by human subjects, it attempts to predict the visual fidelity of 3D meshes according to subjective experiments. This evaluation aims to predict the perceptual visual quality of 3D meshes. However, it is used mainly to validate the performances of objective metrics. In general, the *quality* is the judgment of how two

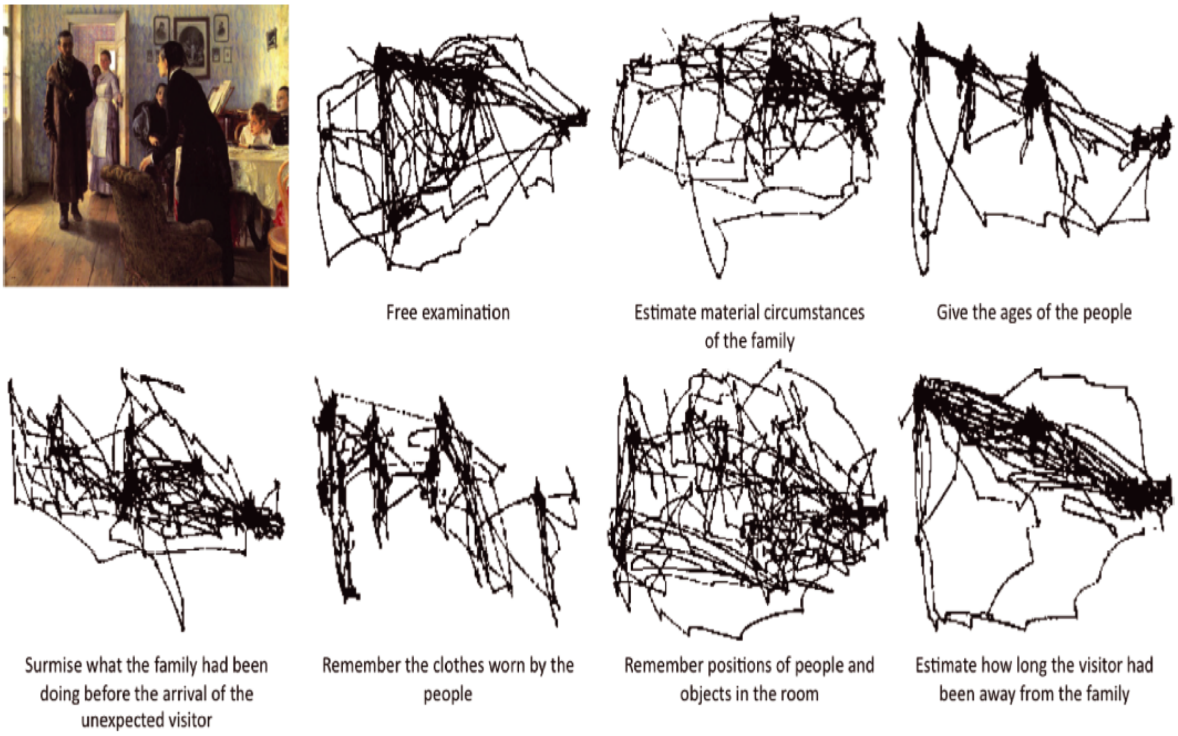


FIGURE 1.4 – Experiments performed by Yarbus Yarbus (2013) on how the task given to a person influences the eye movement.

images, videos, or 3D shapes (the original version is the reference and the other one is distorted or modified) are similar. The results of the subjective evaluation are strongly affected by several parameters used in an experiment, especially for computer-generated stimuli where the user can control the majority of them. The most effective parameters that can be controlled in a subjective quality assessment experiment are listed as follows :

- Lighting : The effect of lighting comes at first importance when talking about viewing conditions. The type of light source and position is crucial for displaying quality. As shown Rogowitz et al. Rogowitz et Rushmeier (2001) objects with lighting from above provide different subjective scores compared to the same objects lif from the front. It has been proven in Howard et Rogers (2002) that the HVS has a prior that light is stationary and comes from a left-above orientation.
- Background : The background can also affect the quality estimation, it changes the visibility of the boundaries of the model. In Corsini *et al.* (2007), authors choose to use a background fading from blue to white, the reason behind this choice is to avoid the overestimation of the contours. Several researches use a uniform black background Watson *et al.* (2001); Rushmeier *et al.* (2000).
- Animation and Interaction : In the subjective evaluation experiments, observers can manipulate the 3D shape and see it from different angles to ensure a fair evaluation. This is done by giving the freedom to observers to control the viewpoint and animating the 3D shape. This parameter inevitably affects the perception of the model. The artifacts caused by simplification are more visible when the object

is not moving, whereas rotating the model makes the distortions less visible. In Daly (2001), the authors claim that the sensitivity of the HVS depends on retinal velocity and the ability of the eye's tracking is limited to 80 deg/sec.

— Type of objects : The type of the used objects in the experiments plays an important role in the evaluation. In Watson *et al.* (2001), authors claim that animal objects and man-made distortions provide different assessment results. Besides, the roughness and the complexity also have considerable influence, that is to say, the artifacts in a rough model are less visible and might be masked.

— Masking : the visual masking is described in details in 1.2.2, the geometric attributes such as roughness, texture, watermarking, and so on, can mask each other. Many works have adopted this effect. In Lavoué (2009) and Pan *et al.* (2005), the authors study the masking effect of roughness and textures on geometry respectively. The masking effect should be considered while designing an experiment.

— Extent : The rendered 3D object in 2D images is displayed in an area called the extent. To reflect the details of the model, this area should be large enough in the experiments.

— Levels : When an operation (simplification, watermarking, etc.) on meshes is to be tested, the number of the comparison cases and the strengths of the applied operations for each case should be adjusted carefully. Too few levels (compared cases) may not sufficiently reflect the tested operation, whereas a large number of levels may not be feasible, as they would require too many subjects. For simplification case, there are studies using three Watson *et al.* (2001), Rogowitz et Rushmeier (2001) to seven Silva *et al.* (2008) levels (including the originals) of simplification.

— Stimuli order : In comparison-based experiments, stimuli can be shown to the user simultaneously (e.g., side-by-side) or in succession (e.g., first the reference, then the tested models). When they are shown in succession, enabling users to turn back to the reference model as in the experiment of Rogowitz and Rushmeier Rogowitz et Rushmeier (2001), allows for a more detailed comparison. Also, the order and the position of the stimuli should be selected in a way that minimizes the effect of external variables such as observer movements and the room's ambient light.

— Duration : The period of showing the models to the observers may also affect the results of the subjective assessment.

1.4 Perceptual objective methods

Objective methods can be classified according to the availability of the reference : full-reference (FR) when the reference model is fully available, reduced-reference (RR) when only a part of the reference is available and no-reference (NR) when no information of the reference is available. In the literature, many approaches have been proposed to evaluate the visual quality of a distorted mesh. Simple geometric measures have been used to evaluate the difference between a reference model and a distorted version such as the root mean squared error RMS Cignoni *et al.* (1998) and the Hausdorff distance Aspert *et al.* (2002) that use a simple similarity between the reference mesh and the distorted one. However, similarly to the mean square error (MSE) and the peak signal-to-noise ratio (PSNR) for the case of digital images Wang et Bovik (2009), it has been demonstrated that this type of metrics generally fails to reflect the perceived visual quality, and do not correlate well with human perception Lavoué et Corsini (2010); Bulbul *et al.* (2011) since it computes a pure geometric distance neglecting the main operations of the HVS. Otherwise, many researchers tend recently to develop perceptually driven quality methods for 3D meshes to take into account the main function of the HVS. It is thus necessary to make use of perceptually relevant features and take into account some important properties of HVS to develop an effective objective MVQ method which correlates well with human perception. It has been demonstrated that incorporating the perceptual properties of the HVS, such as the visual masking effect, will benefit various 3D multimedia applications Lin et Kuo (2011); Lin *et al.* (2012). Several existing quality metrics for 3D meshes Lavoué *et al.* (2006); Lavoué (2011); Wang *et al.* (2012); Torkhani *et al.* (2012) have considered the characteristics of the HVS for a better estimation of the perceived quality. It is now believed that the perceptually driven methods that can effectively predict the subjectively perceived quality of the visual degradation of a 3D mesh will replace the classical geometric measures in a wide range of geometry processing applications in the future. For the no-reference quality assessment, several researchers tend to use machine learning methods which are powerful mathematical tools for prediction problems and had been used in several no-reference methods in the literature for image quality assessment (IQA). To evaluate the effects of the image enhancement filter, a circular back propagation-based image quality method was proposed in Gastaldo et Zunino (2005) and Gastaldo *et al.* (2005). These methods use general pixel-based image features without taking into consideration the perceptual factors. In Babu *et al.* (2007), the authors assess JPEG image artifacts using a growing and pruning radial basis function network. In Chetouani *et al.* (2010), the distortion in blurred images is evaluated using a multilayer perception network to fuse three no reference methods into a single device. In Narwaria et Lin (2010), the image quality is predicted using the support vector regression, in this work the authors tend to use singular vectors out of singular value decomposition as features for quantifying major structural information in images. A new two-step framework called blind image quality index (BIQI) is proposed in Moorthy et Bovik (2010), this method is based on natural scene statistics (NSS) models used for no reference IQA. Another no reference IQA method was proposed in Saad *et al.* (2010), eight NSS features expressed as statistics of local discrete cosine transform (DCT) coefficients are used to predict the image quality. In Li *et al.* (2011a) authors propose an algorithm for no reference IQA which can handle multiple distortions, the proposed algorithm is based on the general regression neural network (GRNN) and uses perceptually relevant features rather than deploying NSS-based features.

1.4.1 Distance based metrics

HD

The Hausdorff distance is one of the most popular and earliest metrics for comparing a reference mesh with its distorted version. This metric calculates the similarity between the two models by computing the following distance :

$$HD(S_0, S_\alpha) = \max_{p_0 \in S_0}(d(p_0, S_\alpha)) \tag{1.1}$$

with

$$d(p_0, S_\alpha) = \min_{p_\alpha \in S_\alpha} ||p_0 - p_\alpha|| \tag{1.2}$$

where S_0 and S_α are respectively a surface of the original and degraded meshes. $d(p_0, S_\alpha)$ represents the distance between a point p_0 to a surface S_0 and a surface S_α.

RMS

The root mean square is another distance-based metric. It calculates the difference between the reference and the distorted meshes to estimate how similar they are. The RMS difference is computed as follows :

$$RMS(M, M') = \sqrt{\sum_{i=1}^{n} ||m_i - m'_i||^2} \tag{1.3}$$

where M and M' are respectively the reference and the distorted meshes, m_i and m'_i are the corresponding vertices of M and M', and $||.||$ is the Euclidean distance between two points.

Similarly to the MSE and the peak signal-to-noise ratio (PSNR) for the case of digital images, it has been demonstrated that this type of metrics generally fails to reflect the perceived visual quality, and do not correlate well with human perception since it computes a pure geometric distance neglecting the main operations of the human visual system (HVS). However, these metrics are relatively well for noise degradation and could be adopted for this type of degradation.

1.4.2 Perceptual metrics

Many researchers tend recently to develop perceptually driven quality methods for 3D meshes to take into account the main function of the HVS.

3DWPM1 and 3DWPM1

The 3DWPM2 method proposed by Galesca et al Corsini *et al.* (2007) is a roughness-based measure. This method has been initially developed for watermarked meshes and it is an improved version of 3DWPM1 Gelasca *et al.* (2005). The authors propose to estimate the roughness on smooth areas by computing the variance of the differences between both original and the watermarked mesh. The roughness-based difference is calculated as follows :

$$R(M_o, M_d) = \log\left(\frac{R(M_o) - R(M_d)}{R(M_o)} + c\right) - \log(c) \tag{1.4}$$

where $R(M_o)$ and $R(M_d)$ is the roughness of the original mesh and the distorted mesh respectively, and c is a constant used to avoid numerical instability.

The first roughness measure (3DWPM1) is a variant of the method by Wu et al. Wu *et al.* (2001). This metric measures the per-face roughness by making statistical considerations about the dihedral angles, i.e. the angle between the normals of two adjacent faces. The idea is that the dihedral angle is related to the surface roughness. The face normals of a smooth surface vary slowly over the surface, consequently, the dihedral angles between adjacent faces are close to zero. To take into account the scale of the roughness the per-face roughness is turned into a per-vertex roughness and rings of different sizes (1-ring, 2-ring, etc.) are considered during roughness evaluation. The total roughness of the 3D object is the sum of the roughnesses of all vertices. The second method by Drelie Gelasca et al.Gelasca *et al.* (2005) (3DWPM2) is based on the consideration that artifacts are better perceived on smooth surfaces. Following this statement, this approach applies a smoothing algorithm and then measures the roughness of the surface as the variance of the differences between the smoothed version of the model and its original version.

MSDM

Inspired by the 2D well-known metric SSIM (Structural Similarity Metric) Wang *et al.* (2004), the MSDM (Mesh Structural Distortion Measure) Lavoué *et al.* (2006) metric is a structural-based method and is achieved by combining three factors : Local curvature (L), Contrast (C) and Structural (S). The local measure is computed as follows :

$$\begin{aligned} LMSDM(x, y) = (\alpha \times L(x, y)^{\alpha} + \\ \beta \times C(x, y)^{\alpha} + \gamma \times S(x, y)^{\alpha})^{\frac{1}{\alpha}} \end{aligned} \tag{1.5}$$

with

$$\begin{aligned} L(x, y) &= \frac{||\mu_x - \mu_y||}{\max(\mu_x, \mu_y)} \\ C(x, y) &= \frac{||\sigma_x - \sigma_y||}{\max(\sigma_x, \sigma_y)} \\ S(x, y) &= \frac{\sigma_x\sigma_y - \sigma_{xy}}{\sigma_x\sigma_y} \end{aligned} \tag{1.6}$$

where x and y indicate respectively two local windows of the original mesh and its degraded version. The global measure is finally given by the following equation :

$$MSDM(X,Y) = \left(\frac{1}{\omega}\sum_{i=1}^{\omega} LMSDM(x_i, y_i)^{\alpha}\right)^{\frac{1}{\alpha}} \tag{1.7}$$

where ω is the number of local windows in the meshes and a is a parameter chosen regarding specific experimental results. The authors propose an improved version of this method in Lavoué (2011) by using a multi-scale approach.

DAME

The DAME metric Váša et Rus (2012) is based on the estimation of Dihedral angles. The idea developed by the author is to measure the variations of these angles for all edges of the mesh. They also integrate a masking weight based on the smoothness and a visibility weight. On each pair of neighboring triangles t_1 and t_2 of the mesh, the oriented dihedral angle is calculated according to the expression :

$$D_{t_1 t_2} = \arccos(n_1, n_2) \times sgn(n_1(s_4 - s_3))$$

where n_1 and n_2 are the normals associated to the triangles t_1 and t_2 respectively.

The perceptual distance measurement DAME is computed as follows :

$$DAME = \frac{1}{n_e}\sum_{ne} ||\alpha_0(i) - \alpha_d(i)||.m_i.w_i \tag{1.8}$$

where n_e represents the number of edges. $\alpha_0(i)$ and $\alpha_d(i)$ are respectively the dihedral angles of the original and degraded meshes. m_i is the masking weight, while w_i is the visibility weight.

GL1

The GL1 method proposed by Karni and Gotsman Karni et Gotsman (2000) is a roughness-based error metric. This measure was used to evaluate their mesh compression approach. It computes the Geometric Laplacian of a vertex v_i as follows :

$$GL(v_i) = v_i - \frac{\sum_{j \in n(i)} l_{ij}^{-1} v_i}{\sum_{j \in n(i)} l_{ij}^{-1}} \tag{1.9}$$

where $n(i)$ is the set of neighbors of vertex i, and l_{ij} is the geometric distance between vertices i and j.

$GL(v)$ represents the difference vector between v and its new position after a Laplacian smoothing step. Considering Fig. 1.9 Karni and Gotsman have derived a visual metric $GL1$ between two objects A and B defined as :

$$GL1(A,B) = \alpha\ RMS(A,B) + (1-\alpha)\left(\sum_{i=1}^{n}\left\|GL(v_i^A) - GL(v_i^B)\right\|^2\right)^{1/2}$$

A developed version called GL2 is proposed in **?** with $\alpha = 0.15$ instead of $\alpha = 0.5$ for GL1.

FMPD

Another perceptual metric called Fast Mesh Perceptual Distance (FMPD) proposed by Wang et al Wang *et al.* (2012) in 2012. It is a reduced reference metric based on the comparison of global roughness measurements computed on two meshes under comparison.

As the authors believed that there is a link between mesh curvature and the perceived mesh quality, the local roughness is defined at each vertex of the reference mesh and the deformed mesh based on Gaussian curvature. First, for each vertex v_i in the reference mesh M_r, the discrete Gaussian curvature is defined as follows :

$$GC_i = \left|2\pi - \sum_{j\in N_i^(F)} \alpha_j\right|$$

where $N_i^(F)$ is the set of all the neighboring facets of v_i, and α_j is the angle in facet j that is incident to v_i.
The local roughness at v_i is measured as the Laplacian of the discrete Gaussian curvature. Thus, compute the mesh Laplacian matrix is first computed as :

$$\begin{cases} D_{i,j} = \frac{\cot(\beta_{i,j})+\cot(\beta'_{i,j}))}{2} \text{ for } j \in N_i^(V)\ , \\ D_{i,i} = -\sum_j D_{i,i} \end{cases}$$

where $N_i^(V)$ is the set of all the neighboring vertices of v_i, $\beta_{i,j}$ and $\beta'_{i,j}$ are the two angles opposite to the edge that connects v_i and v_j. After that, the local roughness LR_i at v_i is defined as :

$$LR_i = \left|GC_i - \frac{\sum_{j\in N_i^(V)} D_{i,j}\cdot GC_j}{\sum_{j\in N_i^(V)} D_{i,j}}\right| = \left|GC_i + \frac{\sum_{j\in N_i^(V)} D_{i,j}\cdot GC_j}{D_{i,i}}\right|$$

The same computation is performed at each vertex v'_i on the deformed mesh M_d. The visual masking and the psychometric saturation effect are considered by a modulation of the local roughness and the values RLF_i for each vertex v_i are obtained. After that, the global roughness using the normalized surface integrals of the local roughness on both meshes is computed.

$$GR = \frac{\sum_i LRF_i \cdot s_i}{\sum_i s_i},$$

where s_i is one-third of the total area of the incident facets of v_i. Finally, the perceptual distance between the reference mesh M_r and the distorted mesh M_d is evaluated as the difference between the two surface integrals.

$$FMPD_{M_r,M_d} = c|GR - GR'|$$

where GR and GR' are the global roughness of M_r and M_d, respectively, and c is a scaling factor.

TPDM

Torkhani et al Torkhani *et al.* (2012) proposed an effective perceptual metric to estimate the perceived quality of distorted triangular meshes. This metric is based on the measurement of a distance between curvature tensors of the two meshes under comparison. It uses the tensor eigenvalues (i.e., curvature amplitudes) as well as the tensor eigenvectors (i.e., principal curvature directions) to derive a perceptually-oriented tensor distance. It also includes the visual masking effect of the human visual system, through a roughness-based weighting of the local tensor distance. A final score that reflects the visual difference between two meshes is obtained via a Minkowski pooling of the weighted local tensor distances over the mesh surface. A local tensor distance is computed for each pair of v_i and $v'_{i,k}$ as :

$$LTD_{v_i,v'_{i,k}} = \frac{\theta_{\min}}{\pi/2}\delta_{k_{\min}} + \frac{\theta_{\max}}{\pi/2}\delta_{k_{\max}}$$

where $\theta_{min} \in [0, 2\pi]$, $\theta_{max} \in [0, 2\pi]$, $\delta_{k_{\min}}$ and $\delta_{k_{\max}}$ are parameters that establish the the correspondence relationship between the curvature amplitudes/directions, which is based on the minimum angular distance criterion between principle curvature directions. In order to consider the visual masking effect, the authors modulate the values of $LTD_{v_i,v'_{i,k}}$ by two roughness-based weights (the rougher the local surface is, the smaller the weights are). The local perceptual distance between v_i and $v'_{i,k}$ which incorporates the visual masking effect, is computed as :

$$LPD_{v_i,v'_{i,k}} = RW_i^{(\gamma)} \cdot RW_i^{(k)} \cdot LTD_{v_i,v'_{i,k}},$$

with $RW_i^{(\gamma)}$,$RW_i^{(k)} \in [0.1, 1.0]$. They are respectively the roughness-based weights derived from surface principal directions and curvature amplitudes in 1-ring neighborhood of v_i. the local TPDM distance associated to v_i, denoted by $LTPDM_{v_i}$ is computed as the barycentric interpolation of the three local perceptual distances, between respectively v_i and $v'_{i,1}$, v_i and $v'_{i,2}$, and finally v_i and $v'_{i,3}$:

$$LTPDM_{v_i} = \sum_{k=1}^{3} b_k(v'_i) LPD_{v_i,v'_{i,k}}$$

The global tensor-based perceptual distance measure TPDM from the reference mesh M_r to the distorted mesh M_r is computed as a weighted Minkowski sum of the local distances $LTPDM_{v_i}$,$i = 1, 2, 3 \ldots, N$:

$$TPDM = \left(\sum_{k=1}^{3} \omega_i |LPTDM_{v_i}|^p \right)^{\frac{1}{p}},$$

where $\omega_i = s_i / \sum_{i=1}^{N}$ with s_i one third of the total area of all the incident facets of v_i, p is a constant that increase the importance of the local distances of high amplitude in the calculation of the global perceptual distance.

Strain Field-based Measure (SF)

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

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

ω_i are the weights and W_i is the strain energy associated with each triangle. Varying the ω_i weights Zhe Bian et al. tested some variants of this metric, but from their experimental results they concluded that the simpler one $(\omega_i = 1)$ gave results similar to the other variants, hence the unweighted one is preferable due to its simplicity.

FR metric based on a support vector regression

In 2018, Chetounai Chetouani (2018a) proposed a full reference method based on extracted features and the support vector regression. The method explores many features proposed in the literature that can be exploited. The author proposes to use directly some existing 3D mesh quality metrics, which are better adapted to estimate the visual quality. After several experimental tests, some geometric attributes were also considered to analyze the quality of 3D meshes. The features were experimentally chosen through several combinations. The tested combinations were performed under the assumption that each selected feature extracts relevant information the selected features are : geometric distance, global roughness (dihedral angle), Global roughness (smoothness), mean of the maximum local curvature directions and the variance of the local curvatures.
Once the features had been selected and extracted, a support vector regression (SVR), is used as the combination tool. the inputs of the SVR model are the extracted features, while its output is the predicted objective score. During the training step, the target is the subjective score (MOS) provided for each mesh.

BMQI

Another perceptual method proposed recently in 2019 by Nouri et al Nouri *et al.* (2017) called Blind Mesh Quality Index (BMQI). Given a distorted mesh, a multiscale saliency map MS and a roughness map R are first computed. Then the mesh is segmented into a number of super-facets N_{SF}. These super-facets play the role of local patches since the human visual system (HVS) locally processes the information. Once the segmentation is performed, each vertex v_i of a super-facet SF_j is labeled by its respective values of saliency $MS(v_i)$ and roughness $R(v_i)$. Then, a feature vector of four attributes for each super-facet SF_j is constructed :

$$\Phi_j = [\mu_{SF_j}, \sigma_{SF_j}, \delta_{SF_j}, \gamma_{SF_j}] \qquad \text{with } j \in [1, N_{SF}]$$

where μ_{SF_j} and σ_{SF_j} represent respectively the local mean saliency and local standard deviation saliency of the superfacet SF_j. These parameters are defined as :

$$\mu_{SF_j} = \frac{1}{|SF_j|} \sum_{v_i \in SF_j} MS(v_i)$$

$$\sigma_j = \sqrt{\frac{1}{|SF_j|} \sum_{v_i \in SF_j} (MS(v_i) - \mu_j)^2}$$

where $|SF_j|$ represents the cardinality (i.e, the number of vertices) of the superfacet SF_j.
δ_{SF_j} and γ_{SF_j} denote respectively the local mean roughness and the local standard deviation roughness that are defined as :

$$\delta_{SF_j} = \frac{1}{|SF_j|} \sum_{v_i \in SF_j} LRF(v_i)$$

$$\sigma_j = \sqrt{\frac{1}{|SF_j|} \sum_{v_i \in SF_j} (LRF(v_i) - \delta_j)^2}$$

The constructed feature vector is used for the learning step. The authors adopt the support vector regression (SVR) that is also used for scoring the visual quality of the 3D mesh.

1.5 Databases and validation protocol

Practically and whatever the type of media (image, video, or 3D models), the design of a subjective test is composed of the following steps :

1. A database is constructed containing different objects (reference objects and distorted versions).
2. The subjective experiment is conducted where human observers give their opinion or some ratings about the perceived distortions of the database objects. A MOS

is then computed for each distorted object of the corpus : $MOS_i = \frac{1}{n}\sum_{n}^{j=1} m_{ij}$, where MOS_i is the MOS of the i^{th} object, n is the number of test subjects and m_{ij} is the score (in a given range) given by the j^{th} subject to the ith object.

3. Because some observers may have used the rating scale differently, a normalization of the MOS values is usually conducted, followed by the filtering of a possible outlier. The reliability of the MOS may also be checked by computing the 95% confidence intervals or the intra-class correlation coefficient.
4. The correlation can then be computed between the MOSs of the objects and their associated metric values ; usually two correlation coefficients are considered : the Spearman Rank Order Correlation and the Pearson Linear Correlation Coefficient. The Pearson correlation is computed after performing a non-linear regression on the metric values, usually using a logistic or a cumulative Gaussian function. This optimizes the matching between the values given by the objective metric and the subjective opinion scores provided by the subjects.

As raised recently by Ebrahimi Ebrahimi (2009), the design of subjective tests producing reliable and reproducible MOS is a delicate task which depends on several ingredients :

- The environment, i.e. type of monitors, viewing distances, lighting conditions.
- The material, i.e. the test objects. The corpus should contain different kinds of models and different types of distortions and not focus on a specific scenario. The range of the visual impacts of the distortions has to be correctly balanced. It is also better to present worst-case models (i.e. anchor conditions) to allow the observers to calibrate their ratings.
- The methodology, i.e. how to present the distorted models and how to rate them. The distorted model can be displayed together with its original version (Simultaneous Double Stimulus) or alone (Single Stimulus). The rating can be categorical adjectival (bad, poor, fair, good, excellent), categorical numerical (1,2,3,4,5) or on a continuous scale (e.g. $\in [0, 100]$).
- The analysis of the data, i.e. how to make sure that the MOS is significant.

1.5.1 Datasets

MVQ goal is to provide quality predictions correlated with the human observer's opinion. To test the performances of MVQ algorithms, a dataset of distorted meshes graded by human observers is needed. As the evaluation process is greatly influenced by the geometric aspect of the meshes, care must be taken when choosing the dataset. It must contain meshes that reflect adequate diversity in their content and generated distortions should reflect a broad range of mesh degradation. To comply with this argument, our approach has been tested and validated using three datasets of distorted and scored meshes (the mean opinion score (MOS) values are provided for each dataset) specially designed for quality metrics evaluation. Fig. 1.5 shows the reference objects that contain the three databases. These latter are defined as follows :

- LIRIS/EPFL General-Purpose database [1] Lavoué *et al.* (2006) : This database was created at the EPFL, Switzerland. It contains 4 reference meshes, Armadillo, Dyno,

[1] http ://liris.cnrs.fr/guillaume.lavoue/data/datasets.html

Venus and RockerArm, and 84 distorted models (88 models total). Two types of distortion are applied, smoothing and noise addition, either locally or globally on the reference mesh. The subjective evaluation was done by 12 observers. The given scores are between 0 (good quality) and 10 (bad quality), and for each model, a normalized MOS is computed by averaging all the scores given by the 12 observers.

- LIRIS Masking database[2] Lavoué (2009) : This database was created at the University of Lyon, France. It contains 4 reference meshes, Armadillo, Bimba, Dyno and Lion, and 24 distorted models (28 models total). The local noise addition is the only type of distortion applied. The specific objective of this database is to test the capability of MVQ methods in capturing the visual masking effect. The subjective evaluation was done by 11 observers. A normalized score between 0 (bad quality) and 4 (good quality) is assigned to each distorted model.
- UWB compression database[3] Váša et Rus (2012) : includes 5 reference models, and 63 distorted models (68 models total), for each reference model twelve or thirteen distorted versions, are created. The reference models are subject to thirteen types of compression distortions. The distorted models of each reference model are organized and sorted according to the perceived quality. A score between 0 (good quality) and 1 (bad quality) is assigned for each distorted model.
- The IEETA simplification database Silva *et al.* (2009) : This database contains five reference models and 30 simplified versions (six distorted versions for each reference). The simplified models were obtained using three simplification algorithms with two different vertex reduction ratios.

1.5.2 Validation protocol

The correlation between the perceptual scores produced by the method and the MOS provided by subjects is used as criteria to evaluate the performance of an objective MVQ method. Usually, two types of correlation coefficients are commonly used i.e. the Pearson linear correlation coefficient (r_p) which employed to measure the prediction accuracy and the Spearman rank-order correlation coefficient (r_s) which employed to measure the prediction monotonicity Wang et Bovik (2006). The Spearman rank-order correlation coefficient (r_p) depends only on the rank of the objective scores of the models ; if the rank of objective scores is similar to the rank of MOSs, a high value will be obtained, regardless of the distance between the objective score and the corresponding MOS. For both criteria, a higher value indicates better prediction performance. These measures are defined as follows :

$$r_p = \frac{\sum_{i=1}^{n}(Qs_i - \bar{Q}s)(MOS_i - M\bar{O}S)}{\sqrt{\sum_{i=1}^{n}(Qs_i - \bar{Q}s)^2}\sqrt{\sum_{i=1}^{n}(MOS_i - M\bar{O}S)^2}} \tag{1.10}$$

$$r_s = 1 - \frac{\sum_{i=1}^{n}(rank(MOS_i) - rank(Qs_i))^2}{n(n^2-1)} \tag{1.11}$$

[2] http ://liris.cnrs.fr/guillaume.lavoue/data/datasets.html

[3] http ://compression.kiv.zcu.cz/

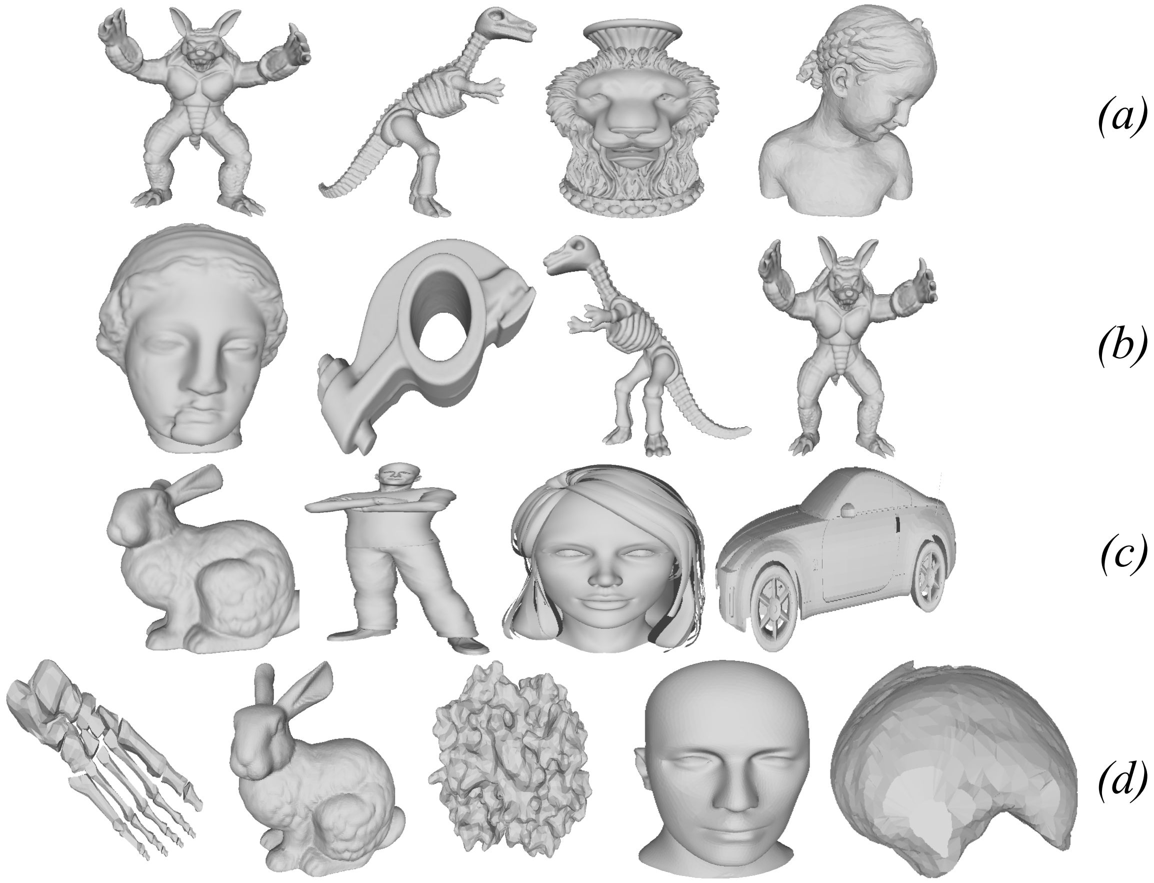


FIGURE 1.5 – The reference models from the LIRIS masking database (a), the general-purpose database (b) and the UWB compression database (c)and the IEETA simplification database (d).

Where n denotes the numbers of distortions in a given database. The mean opinion scores provided in the database are defined by MOS_i, and Qs_i presents the objective quality score obtained by the proposed method.

The values of the scores obtained by the objective method and the mean opinion scores are non-linear, and they are not easy to interpret by users. To partially remove the non-linearity between the objective scores and the MOS, it is highly recommended to introduce a psychometric fitting. Similarly to Wang *et al.* (2012); Torkhani *et al.* (2012), we use a cumulative Gaussian psychometric function Engeldrum (2000) defined by :

$$p(a,b,X) = \frac{1}{\sqrt{2\pi}} \int_{a+bX}^{\propto} \exp - \left(\frac{t^2}{2}\right) dt \tag{1.12}$$

Where X is the objective score, a and b are two parameters to be determined. The two parameters a and b are estimated for each database using the objective values of the distorted versions and the corresponding MOS with the help of the curve fitting toolbox under Matlab.

1.6 Conclusion

The human visual system is a very complex system in which several sensory and cognitive elements interact. The perceptual mechanisms of this system can be used for predicting visual quality, to control or evaluate the processing algorithms, but it also seems relevant to integrate them directly within these processing algorithms. The development of objective measures to estimate the perceptual quality of 3D meshes must take into account these elements. In this context, we presented in this chapter the main characteristics of the HVS to better understand the major phenomenon. Before describing the existing state of the art objective methods, we presented in this chapter the subjective evaluation. This latter is guided by several controlled parameters and serves to validate the performances of objective metrics. This latter relies on the computation of a quality metric to mimic an ideal human observer. It is a good solution to automatically assess the perceived visual quality of a distorted mesh. However, it must correlate well with the subjective assessment process. Several metrics have been developed to measure the perceived quality of distorted 3D meshes. These quality metrics have been evaluated on several subjective corpora and have a strong correlation with the scores delivered by human observers. Almost all existing methods are whether full or reduced reference approaches. They require inevitably the presence of the reference mesh which is not available in most practical situations. Thus, the development of a no-reference MVQ assessment method becomes crucial to remedy this problem. This solution has recently attracted the attention of many researchers in the field of visual quality assessment. It is in this context that in the next chapter we will present a no-reference approach to predict the perceived visual quality of distorted 3D meshes. This method uses a machine learning method and handcrafted feature. Besides, we take advantage of the function of the HVS by including a visual masking module.

Chapitre

2 Quality assessment based on hand-crafted features and machine learning

Sommaire

2.1 Introduction

As indicated in the last chapter, metrics based on simple geometric distances between two meshes (an original mesh and a degraded mesh) are relatively well for noise degradation and could be adopted for this type of degradation. However, they do not allow us to effectively predict the perceived quality of 3D meshes. These metrics compute a pure geometric distance and neglect the limitations of HVS and its specificities. It is thus necessary to make use of perceptually relevant features and take into account some important properties of the HVS to develop an effective objective MVQ. Several existing quality metrics for 3D meshes are also discussed which consider the characteristics of the HVS for a better estimation of the perceived quality. It is now believed that the perceptually driven methods that can effectively predict the subjectively perceived quality of the visual degradation of a 3D mesh will replace the classical geometric measures in a wide range of geometry processing applications. The blind quality assessment has been successfully adopted for image and video processing, many methods in this field of study have shown notable results in terms of the correlation with human judgment. While the studies in mesh quality assessment tend more to adopt full-reference and reduced -reference approaches. However, it is required to address more importance to blind MVQ assessment. In this context, we propose in this chapter a novel method for blind MVQ assessment in order to handle the practical situations where the reference mesh is not available. Our method is based on the hand-crafted features representing different aspects of 3D meshes extracted from the distorted mesh and Machine learning techniques to learn the extracted features and estimate the perceived quality score. In addition, we take into consideration the main function of the HVS by including a visual masking module. In the first section of this chapter, we introduce two perceptual features used to describe the 3D mesh, we adopt in our methods the mean curvature which represent the visual aspect, and the dihedral angles representing the structural aspect of the 3D mesh. In the second section, we describe the different components of our mesh visual quality assessment method based on dihedral angles features and the support vector regression (SVR). In the same vein, we present in the third section another method based on mean curvature features and the general regression neural network (GRNN) for the quality estimation. In the fourth section, we evaluate the performance of our perceptual objective measures on four subjective databases publicly available.

2.2 Hand-crafted features from the 3D mesh

2.2.1 Dihedral angles extraction

The dihedral angle is an important perceptual feature in mesh processing. it is related to the surface roughness and represent the structural information of the 3D mesh. The face normals of a smooth surface vary slowly over the surface, consequently, the dihedral angles between adjacent faces are close to zero.
We denote $t_1 = \{v_1, v_2, v_3\}$ and $t_2 = \{v_1', v_2, v_3\}$ two adjacent triangular faces, N_1 and N_2 are respectively the corresponding normals of t_1 and t_2 as shown in Fig. 2.1 (a).

The dihedral angle Φ by definition is the angle formed by two normals N_1 and N_2 of two adjacent triangular faces. It is calculated by :

$$\Phi = acos\left(\frac{N_1 \cdot N_2}{norm(N_1) \times norm(N_2)}\right) \tag{2.1}$$

where $norm(.)$ is the norm of the vector.

For regular meshes with triangular faces, each primitive is surrounded by 12 triangular face neighborhoods divided by 3 different types as shown in Fig. 2.1(b). Considering the current primitive (c), we have 3 faces sharing an edge with c (white triangles), 3 others sharing a vertex with c by an acute angle (gray triangles), and finally, 6 primitives sharing a vertex with c and form with it an angle of 120° (black triangles). Considering this neighborhood, we obtain for each triangular face a set of 12 dihedral angles. The overall angles vector for a given mesh is then obtained by concatenating all angles computed from the whole 3D mesh.

$$\Phi_i = [\Phi_1, \Phi_2, \Phi_3, ..., \Phi_n] \tag{2.2}$$

where $n = 12 \times N$ is the length of the feature vector, and N is the number of triangular faces in the mesh.

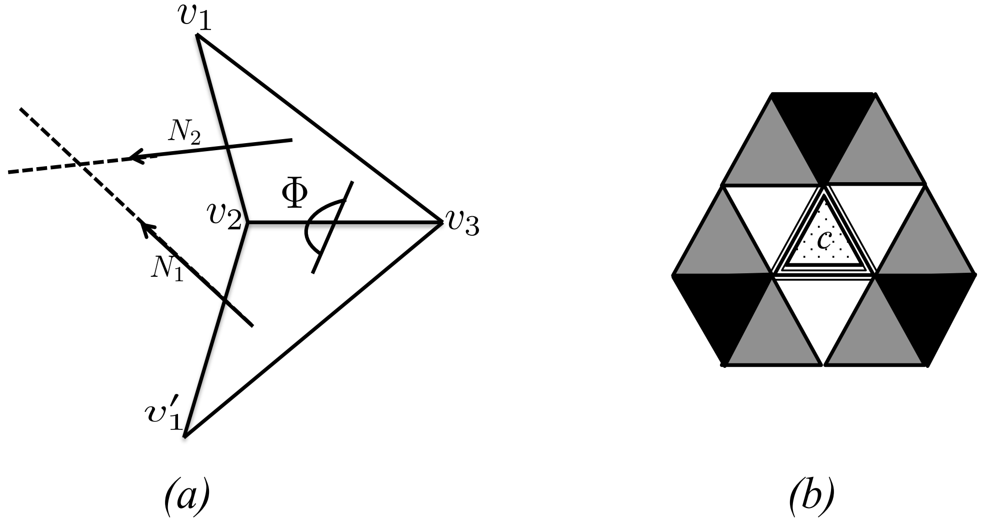


Figure 2.1 – Dihedral angle computation. (a) The dihedral angle between two triangular faces. (b) Illustration of a neighborhood for a regular mesh with triangular faces.

2.2.2 Mean curvature extraction

Due to its ability to describe the visual characteristics of a 3D model, particularly sharpness, roughness or smoothness of a region, curvature features have been adopted in numerous 3D mesh visual quality assessment methods Lavoué (2011); Torkhani *et al.* (2014); Wang *et al.* (2012). Curvature describes the deviation amount of the surface of being flat. Fig. 2.2 shows two mean curvature maps for the Venus model and its distorted version. Fig. 2.2(a) and Fig. 2.2(c) represent respectively the Venus model and the distorted version obtained by applying noise addition on the whole model. After the distortion, the surface of the Venus model becomes rougher, and the mean curvature becomes noisier as it is shown in fig. 2.2(b) and fig. 2.2(d) respectively.

Similarly to Cohen-Steiner and Morvan Cohen-Steiner et Morvan (2003), the curvature tensor denoted T_{v_i} is first calculated at each vertex v_i of the distorted mesh on a geodesic disk neighborhood, defined by the projection of a sphere centered on the vertex

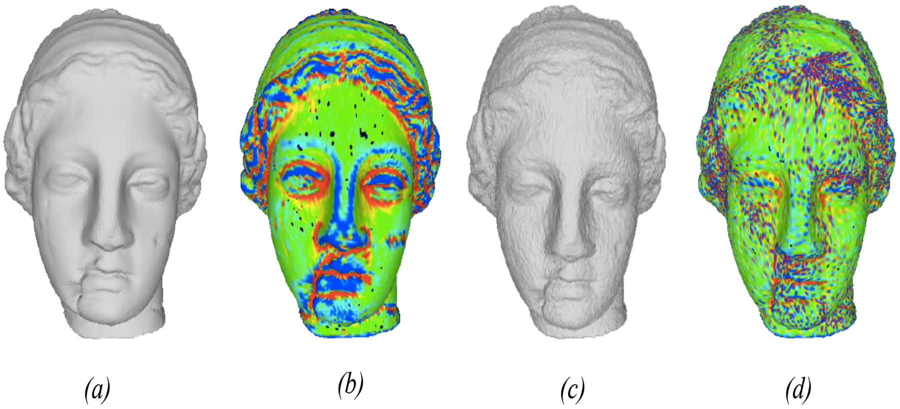


FIGURE 2.2 – Mean curvature describes the visual characteristics of 3D models. (a) Reference Venus model. (b) Mean curvature map of the reference Venus model. (c) A distorted Venus model after noise addition. (d) Mean curvature map of the distorted Venus model.

and the surface of the mesh. For each vertex, the curvature tensor is thus calculated by the following equation :

$$\tau(v) = \frac{1}{|B|} \sum_{egese} \beta(e)|e \cap B|\bar{e}\bar{e}^t, \tag{2.3}$$

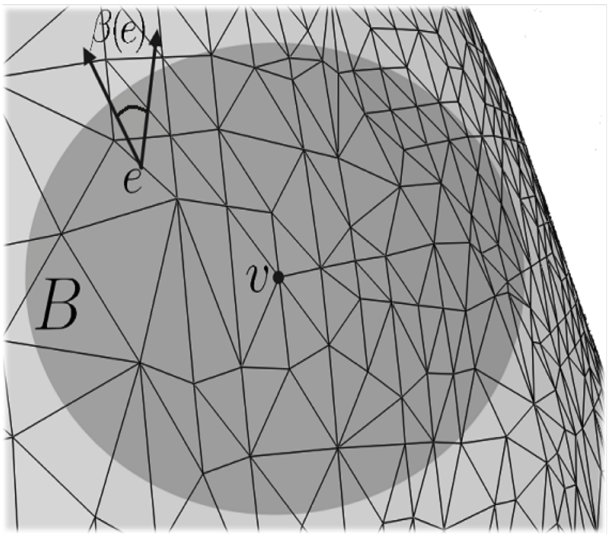


FIGURE 2.3 – Geometric elements used to compute the curvature tensor Torkhani *et al.* (2014)

where B measures the surface occupied by the geodesic disk, $|e \cap B|$ is the length of the edge e located in B. $\bar{e}$ and $\bar{e}^t$ are respectively the vector unit in the direction of e, and

its transpose. $\beta(e)$ are the angles between the normals of the incident facet to each edge e in a geodesic disk B.
Fig. 2.5 shows these geometric elements used to calculate the curvature tensor.

After that, the maximum curvature amplitude k_{max} and the minimum curvature amplitude k_{min} is calculated from the curvature tensor as the absolute values of the two non-zero eigenvalues of the tensor. Then the mean curvature k_{mean} is computed as $(k_{max} + k_{min}/2)$.

2.3 Blind mesh visual quality assessment using dihedral angle features and support vector regression

2.3.1 Pipeline of the method

An overview of the proposed no-reference MVQ assessment method is depicted in Fig. 2.4. The proposed method is no-reference and thus it focuses only on the distorted mesh to predict the quality score without having the reference mesh. As can be seen in Fig. 2.4, our method relies on two main stages (separated by a dashed line in bold) that are training and quality score prediction. In the first stage, given a distorted mesh, we extract the dihedral angles performed by normals of adjacent triangular faces as perceptual relevant features. Afterward, a visual masking modulation is applied on the angle vectors in order to take into consideration the masking effect Breitmeyer (2007). The latter presents an important characteristic of the HVS. Then, we estimate three sets of statistical parameters from the obtained angles vectors using respectively three statistical distributions : Gamma, Weibull and Rayleigh distribution. This step is very crucial owing to its capability to reduce the amount of data to learn, and hence, to optimize the time complexity. The three estimated sets are used to feed the support vector regression is order to construct three regression models. The second stage is dedicated to the quality scores estimation by regression. For doing so, three intermediate objective quality scores (S_G, S_W and S_R) for each set of parameters are determined. Finally, the overall objective quality score Q_S is computed as a weighted sum of the intermediate quality scores by assigned convenient weights ω_g, ω_w and ω_r.

2.3.2 Visual masking effect modulation

To take into consideration the visual masking effect, the visibility of the distortion in rough and smooth regions should not be considered with the same degree. Using dihedral angles as perceptual features, the roughness corresponds to the angles with high magnitude. To reduce the high magnitude, we multiply the angles vector Φ_i by a roughness weight function RW_i defined as follows :

$$RW_i = \exp\left(-\frac{\Phi_i}{2 \cdot \sigma^2}\right) \tag{2.4}$$

where σ is the standard deviation of the dihedral angles vector. The visual masking

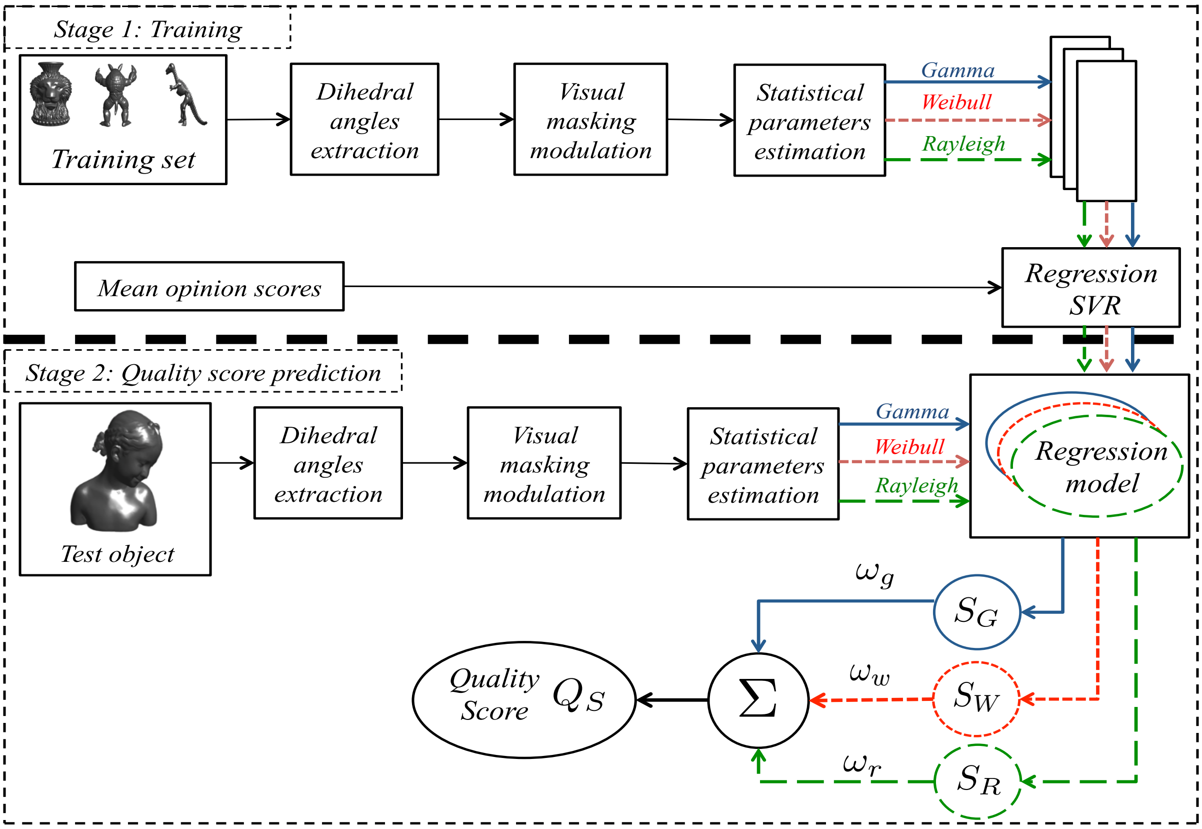

Stage 1: Training
Training set
Dihedral angles extraction
Visual masking modulation
Statistical parameters estimation
Gamma
Weibull
Rayleigh
Mean opinion scores
Regression SVR
Stage 2: Quality score prediction
Test object
Dihedral angles extraction
Visual masking modulation
Statistical parameters estimation
Gamma
Weibull
Rayleigh
Regression model
ω_g
S_G
ω_w
S_W
ω_r
S_R
Σ
Quality Score Q_S

FIGURE 2.4 – Blind mesh quality assessment based on dihedral angles and support vector regression

modulation is then calculated by :

$$\Phi_{masking} = \Phi_i \cdot RW_i \tag{2.5}$$

where $\Phi_{masking}$ is the modulated feature vector. With this module, the visibility of distortions in rough regions will be masked, whereas distortions in smooth regions will not be significantly modified.

2.3.3 Statistical parameters estimation

As mentioned earlier, we adopt the SVR for the feature learning step (see Section 2.3.4 for more details). SVR is a powerful tool for machine learning and pattern recognition that proved its forcefulness in several applications. However, for large-scale problems, SVR is a time and memory consuming due to the enormous resource requirements to solve a quadratic programming (QP) problem. Conversely, a small subset of the large training data called support vectors can be sufficient and adequate to fully determine the SVR decision function. For this reason, it is advisable to restrict the amount of the learning data and only preserve the relevant data for the final decision function.

To remedy this problem, we propose to adopt statistical distributions, namely, Gamma, Weibull, and Rayleigh distributions. The estimation of these distribution parameters allows us to reduce the huge learning sets to small subsets. Fig. 2.5 shows that the estimated models fit well the empirical dihedral angles distribution. Therefore, we can quite simply use the estimated parameters of the models instead of the large training data, and consequently, optimize the computational time.

Before describing these distributions, we recall that the parameters are estimated using the maximum likelihood (ML) method.

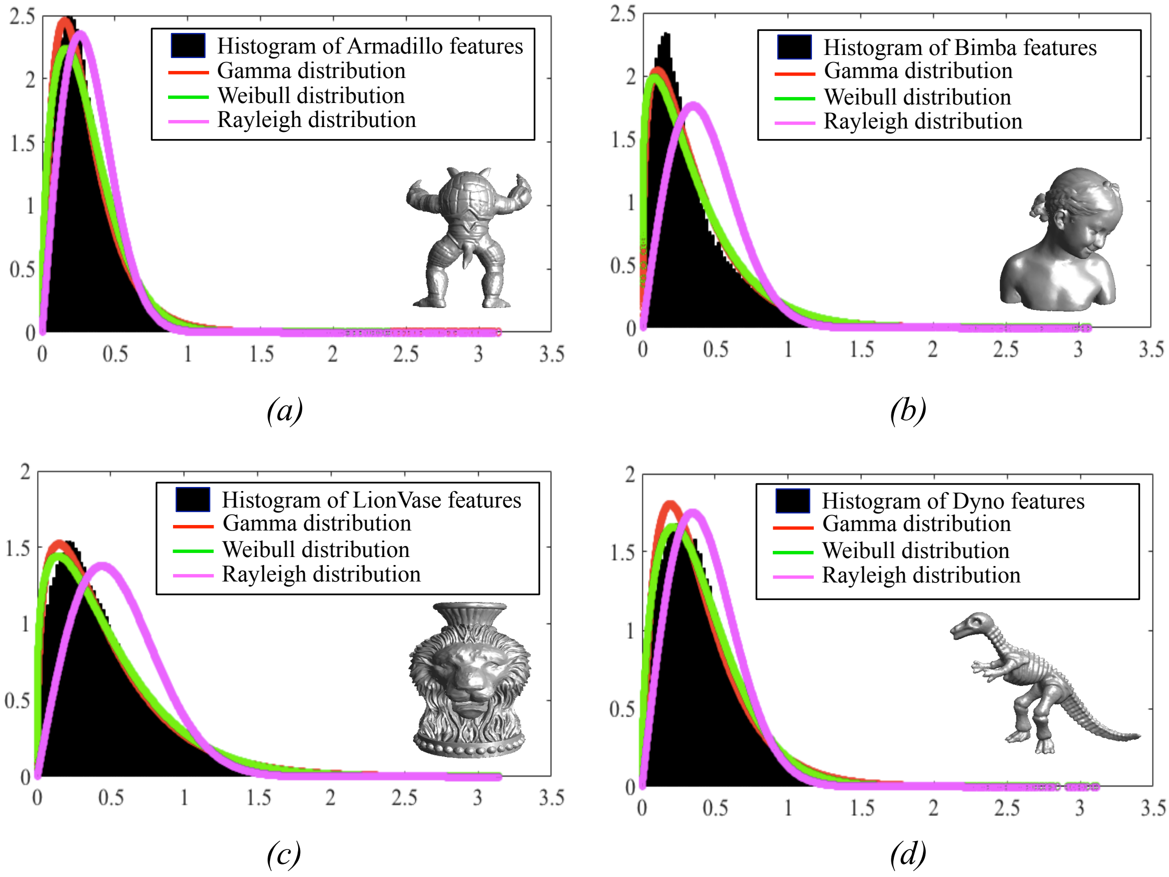


FIGURE 2.5 – Histogram of the extracted dihedral angles from (a) Armadillo, (b) Bimba, (c) LionVase and (d) Dinosaur with noise addition, with their corresponding plots of the fitted Gamma, Weibull, and Rayleigh distribution.

Rayleigh Distribution

Rayleigh distribution presents the most simple choice to describe the extracted features. This distribution is governed by one parameter (shape parameter σ). The probability density function of the Rayleigh distribution is given by :

$$p(x;\sigma) = \frac{x}{\sigma^2}\exp\left(-\frac{x}{2\sigma^2}\right) \qquad 0 < x < \infty \tag{2.6}$$

While $\sigma > 0$ is the shape parameter of the Rayleigh distribution.

Weibull Distribution

Weibull distribution is used to overcome some limitations presented by the Rayleigh distribution modeling. It allows a greater degree of freedom since it is governed by two parameters (shape parameter and scale parameter). The Weibull probability density function is given by :

$$p(x; \tau, \mu) = \frac{\tau}{\mu}(\frac{x}{\mu})^{\tau-1} \exp\{-(\frac{x}{\mu})^{\tau}\} \quad 0 < x < \infty \tag{2.7}$$

While $\tau > 0$ and $\mu > 0$ are respectively the shape parameter and the scale parameter.

Gamma Distribution

Gamma distribution can also replace Rayleigh distribution to model positive data. We say that a random variable x follows Gamma law with a shape parameter b and a scale parameter θ if it admits for probability density :

$$p(x; b, \theta) = \frac{\theta^{-b} x^{b-1}}{\Gamma(b)} \exp\left(-\frac{x}{\theta}\right) \quad 0 < x < \infty \tag{2.8}$$

Where $\Gamma(.)$ denotes the Gamma function.

The estimated parameters of the three presented distributions are then employed as input feature vectors for the support vector regression that will be used for the feature learning and the quality scores estimation.

2.3.4 Feature learning : Support Vector Regression (SVR)

The SVR is an extension of the support vector machines (SVM) Vapnik (2013) for a numeric prediction. SVM is a supervised classification system that finds the maximum margin hyper-plane separating two classes of data. The training instances that are close to this hyper-plane are called support vectors. SVR also produces a decision boundary that can be expressed in terms of a few support vectors and can be used with kernel functions to create complex nonlinear decision boundaries.

In this work, the support vector regression is adopted for the feature learning step to predict the objective quality scores of the distorted meshes. We denote x_i the feature vector for a distorted mesh M_d with a subjective score y_i. The regression function used to estimate an observation x can be expressed as follows :

$$f_{svr}(x) = \sum_{x_i \in V_s} \alpha_i y_i K(x_i, x) + b \tag{2.9}$$

where V_S is the set of support vectors, (x_i, y_i) presents the training set and α denotes the Lagrange multipliers obtained by the minimization process. Furthermore, $K(x_i, x)$ is the kernel function. In this paper, we propose to compare the quality score prediction using four different kernels that are linear, polynomial, radial basis function (RBF), and sigmoid. Table 2.1 shows the different kernels used with their mathematical equations and parameters.

Tableau 2.1 – Kernels used for the support vectors regression.

Kernel	Equation	Parameters
Linear	$K(x_i, x_j) = x_i^T x_j$	-
Polynomial	$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$	γ, d, r
Radial basis function (RBF)	$K(x_i, x_j) = \exp(\gamma \|x_i - x_j\|^2)$	γ
Sigmoid	$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$	γ , r

We note that besides the kernel parameters cited above, SVR involves another parameter C called the penalty parameter of the error term.

2.3.5 Objective quality score prediction

For each set of the estimated statistical parameters, the SVR is used to predict an intermediate quality score conducting a leave-one-out cross-validation (LOOCV) according to the following process :

— We build a training regression model using all the existing 3D objects in the repository except one object and its distorted versions.

— The excluded subset is then used for the test using the constructed regression model.

— The process is repeated for each 3D object in the repository.

Our scheme provides three intermediate scores S_G, S_W and S_R by considering the estimated parameters from Gamma, Weibull, and Rayleigh distributions as inputs for the support vector regression.

Finally, these intermediate quality scores are used to compute the overall quality score by a weighted sum as follows :

$$Q_s = \frac{1}{3}\left(\omega_g S_G + \omega_w S_W + \omega_r S_R\right) \tag{2.10}$$

Where ω_g, ω_w, and ω_r are the weights that the values are fixed in relation to the goodness-of-fit test. More information about the estimation of the weights is given in Section. 2.5.

In order to estimate the weights of intermediate quality scores ω_g, ω_w and ω_r defined by Eq. 2.10, we conduct a Kolmogorov-Smirnov (KS) goodness-of-fit test to find out which statistical distribution describes perfectly the empirical data, and hence to set a convenient weight for each intermediate score. Fig. 2.6 shows the KS measures corresponding to 4 meshes with 6 distorted versions from the LIRIS masking database. Similarly, we conduct the same test for the other databases. Table 2.2 shows the KS test values for each database as well as the average value. We note that a lower KS value indicates a good fitting. In general, Gamma and Weibull distributions provide similar fitting. In addition, they are

Tableau 2.2 – Kolmogorov-Smirnov test measurement for the three databases and the average KS value.

	LIRIS masking	General-purpose	Compression	Average
Gamma	0.25	0.31	0.25	0.27
Weibull	0.30	0.28	0.27	0.28
Rayleigh	0.44	0.56	0.59	0.53

generally better than the Rayleigh distribution, due to the fact that this latter provides only one parameter which is not sufficient to describe the empirical angles data.
According to the average KS values, the weights ω_g, ω_w and ω_r are computed as follows :

$$\omega_g = \frac{1 - KS_g}{\sum_i (1 - KS_i)}, \quad i = \{g, w, r\} \tag{2.11}$$

$$\omega_w = \frac{1 - KS_w}{\sum_i (1 - KS_i)}, \quad i = \{g, w, r\} \tag{2.12}$$

$$\omega_r = \frac{1 - KS_r}{\sum_i (1 - KS_i)}, \quad i = \{g, w, r\} \tag{2.13}$$

where KS_g, KS_w, and KS_r denote the average KS test values for Gamma, Weibull, and Rayleigh respectively.

2.4 Blind mesh visual quality assessment using curvature features and general regression neural network

2.4.1 Pipeline of the method

This method whose the pipeline is delineated in Fig. 2.7, relies on the curvature features extracted from the distorted mesh and the general regression neural network. Our algorithm is a no-reference scheme : we only need the distorted mesh to predict the quality score, while the reference mesh is not involved in any step of the processing. The proposed method is divided into two major steps : perceptual feature extraction and feature learning. The first step of the processing consists of extracting the curvature feature from the distorted mesh as a perceptual feature. that represents the visual characteristics of a 3D model. In the second step, we use the general regression neural network for feature learning to predict the objective quality score.

2.4.2 Feature Learning : General regression neural network

The general regression neural network is one of the most powerful regression tools that fall into the category of probabilistic neural networks that has a dynamic network structure Specht *et al.* (1991); Chartier *et al.* (2009). With only a few training samples

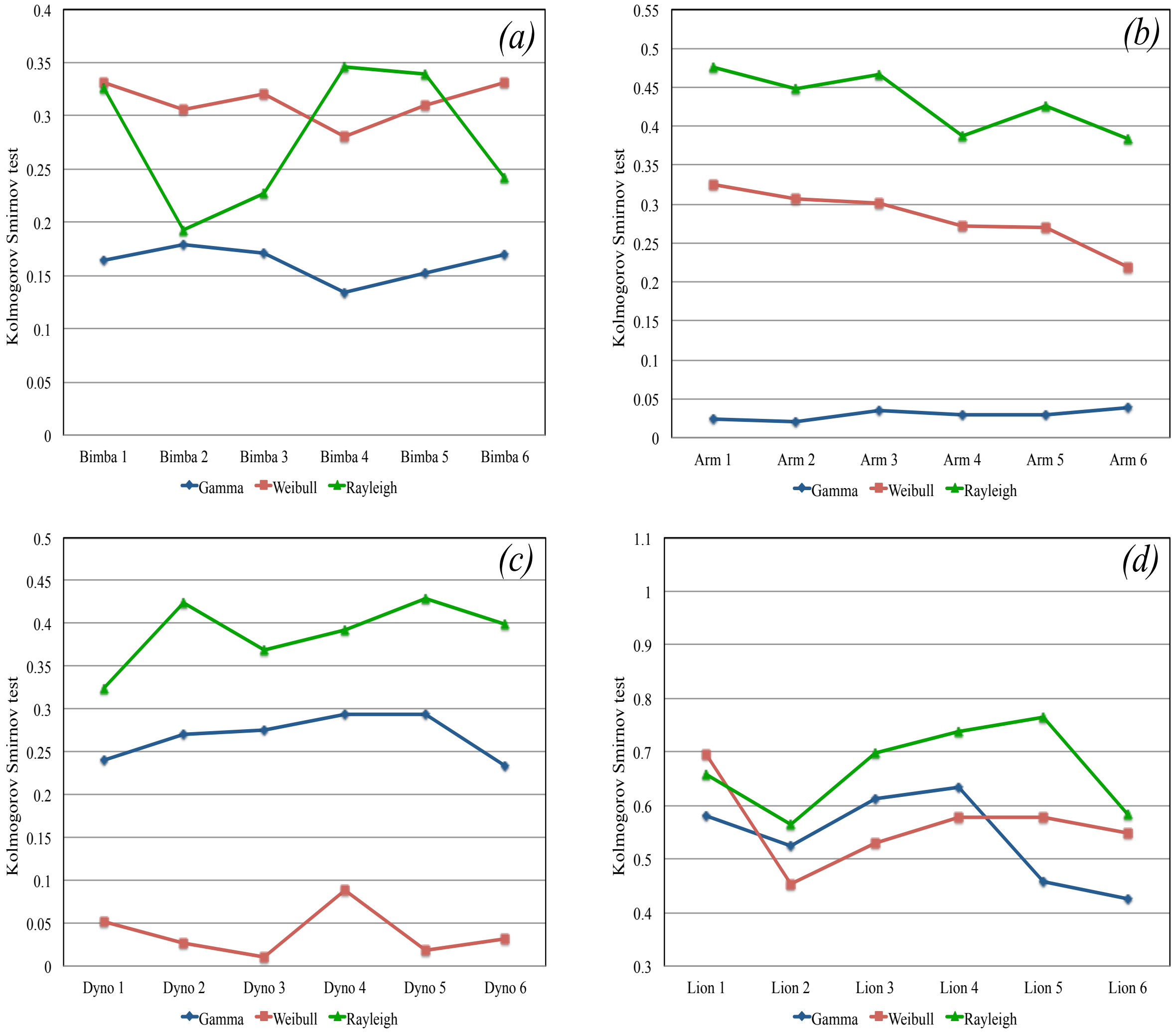


FIGURE 2.6 – Kolmogorov-Smirnov goodness-of-fit test for (a) Bimba model, (b) Armadillo model, (c) Dynosaur model and (d) LionVase model from the LIRIS masking database.

available, GRNN can converge to the underlying function of the data. Due to the fact the available data is generally not enough, it becomes a very useful tool for regression and it has been observed to yield better results than the backpropagation network or RBF network in terms of predictions and comparisons of system performance in practice Li *et al.* (2011b).

GRNN uses the normal distribution as a probability density function. Each training sample is used as the mean of a normal distribution for an input vector X, the output $\hat{Y}$ of the GRNN is obtained as follow Specht *et al.* (1991) :

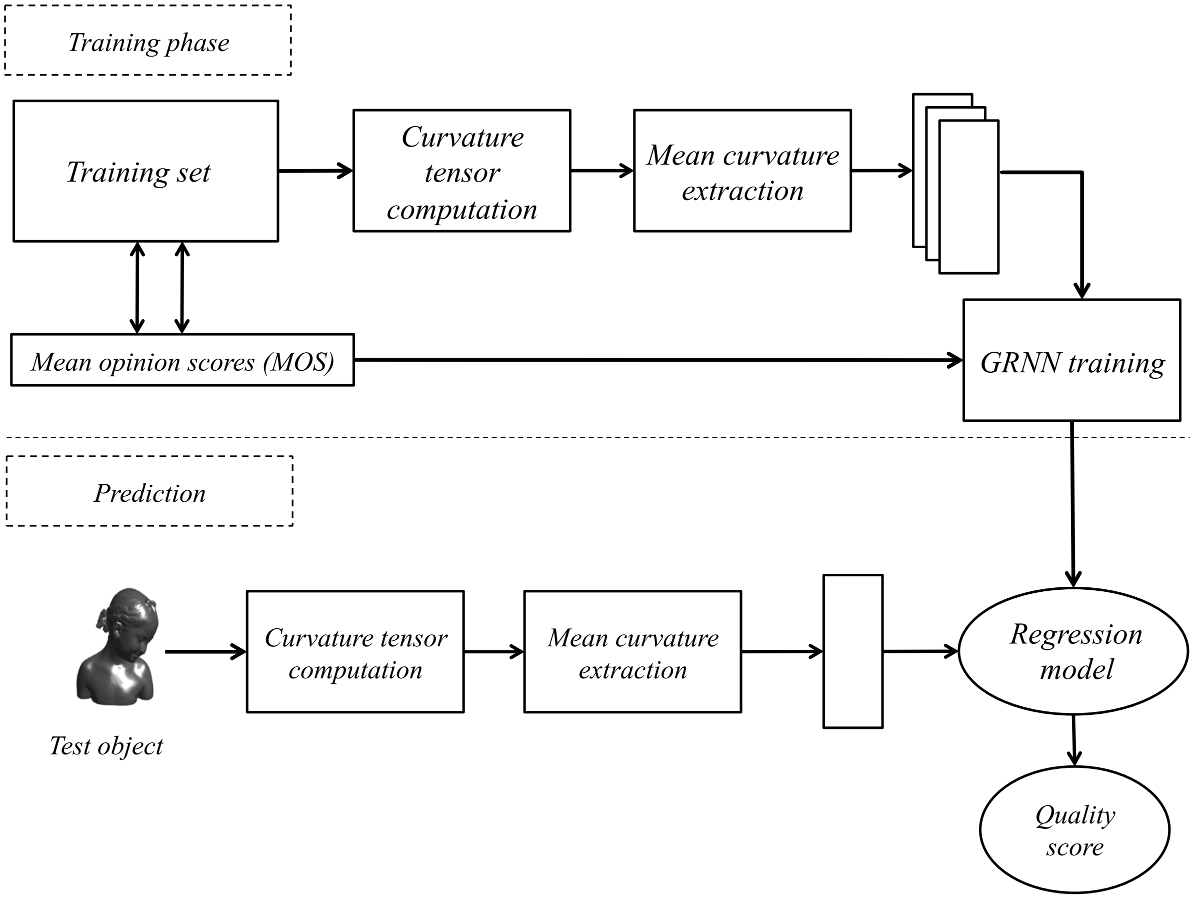

Training phase
Training set
Curvature tensor computation
Mean curvature extraction
Mean opinion scores (MOS)
GRNN training
Prediction
Curvature tensor computation
Mean curvature extraction
Regression model
Test object
Quality score

Figure 2.7 – Blind mesh visual quality assessment using curvature features and general regression neural network.

$$\hat{Y} = \frac{\sum_i^n Y_i \exp(-D_i^2/2\sigma^2)}{\sum_i^n \exp(-D_i^2/2\sigma^2)}, \tag{2.14}$$

where n is the number of sample observations ; $D_i^2 = (X - X_i)^T(X - X_i)$ is the distance between the training sample and the point of prediction used as a measure of how well each training sample can represent the position of prediction ; X_i and Y_i are sample values, and σ is the standard deviation or the smoothness parameter. The larger the value of σ, the smoother the functional approximation. In order to fit the data very closely, the selected value of the smoothness parameter should be smaller than the average distance between the input vectors. Fig. 2.8 delineate a schematic diagram of GRNN for mesh visual quality assessment.

The network consists of four layers : the input layer in which the number of inputs is equal to the number of independent features, in our method we use only the discrete curvature as features extracted from the distorted mesh, thus, the architecture consists only of one input node. The second layer is the pattern layer in which each unit represents a training pattern. The third layer is the summation layer, it includes two-unit, the first

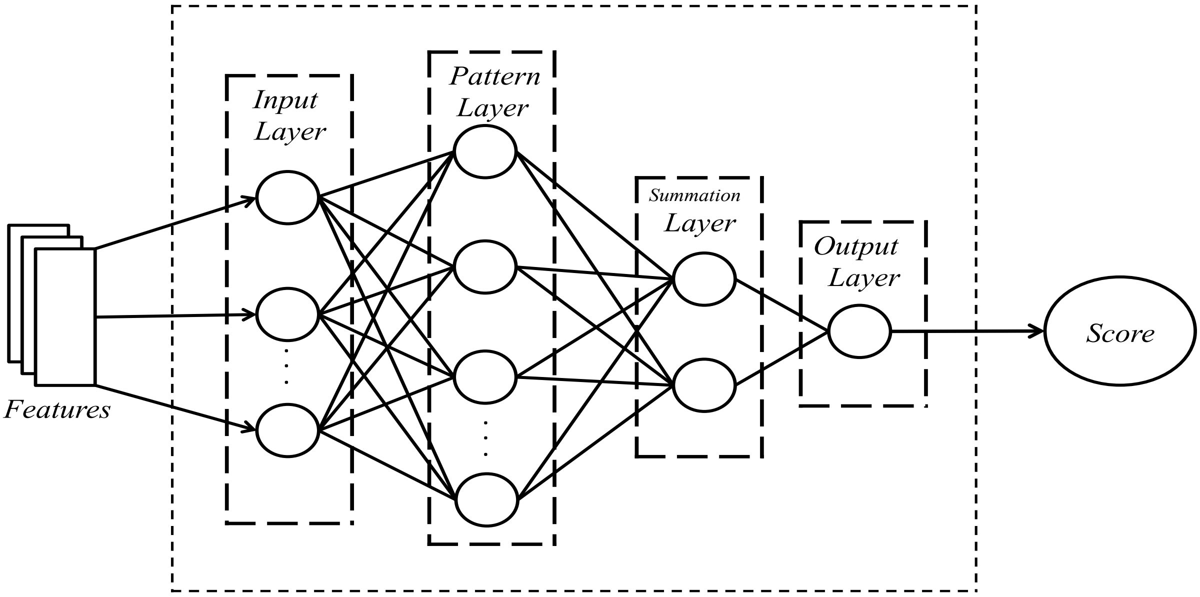


FIGURE 2.8 – Schematic diagram of GRNN for mesh visual quality assessment.

unit assesses the numerator of (2) by summing all the outputs of the pattern layer, while the second unit assesses the denominator of (2). The output layer is the last layer in the general regression neural network architecture, it computes the quotient of the two outputs of the summation layer, in order to predict the final mesh visual quality score. The network weights are crucial parameters in the architecture. The weight of the connection between any node i in the pattern layer and the second node in the summation layer is equal to unity, while the weight on the connection between node i in the pattern layer and the first node of the summation layer is equal to Y_i.
For the training process, we conduct a leave-one-out cross-validation (LOOCV) according to the following process :

- We build a training model using all the existing objects in the repository except one object.
- The excluded object is then used as a test object for the constructed model.
- The process is repeated for each mesh : exclude one object, build a training model, and finally use the excluded object for the test.
- The obtained scores are then compared with the subjective scores using correlation measurements.

2.5 Performances evaluation of the proposed measures

In this section, we evaluate the effectiveness and forcefulness of the proposed no-reference quality assessment methods based on hand-crafted features and machine learning. First, we conduct exhaustive tests on the SVR parameters to retrieve the best system combination that provides the highest performance for the MVQ assessment task. After that, we evaluate the performance of the GRNN and explore its capability to estimate the perceived quality score. Finally, we conduct a comparative study of the proposed methods with the state of the art including FR, RR, and NR methods.

2.5.1 SVR Kernel's parameters selection

The selection of the values of parameters pair (C, γ) is very crucial and strongly affects the prediction results. However, it is not evident to decide which parameter value to take. Therefore, a technique of parameter selection must be introduced. We suggest a grid search on the parameters using the leave-one-out cross-validation Hsu *et al.* (2003). For instance, in the case of RBF kernel, various pairs of (C, γ) values are tested and the best pair is retrieved according to the following process :

- First, we define a grid search space for (C, γ) pair ($C = 2^{(-5:15)}$ and $\gamma = 2^{(-14:2)}$)
- Then, for each (C, γ) pair, we conduct a Leave-One-Out cross-validation on the training set in order to retrieve the optimal pair that leads to the higher accuracy.
- Finally, the regression model is created using the optimal parameter values.

Fig. 2.9 shows an example of a coarse grid search on $C = 2^{(-5:15)}$ and $\gamma = 2^{(-14:2)}$ using the radial basis (RBF) kernel and the LIRIS masking database features. The cross-validation rate for a pair of parameters (C, γ) is highlighted by colors from blue to yellow corresponding respectively to the rates 0.1 and 0.8. The better pair of parameters is the one that leads to a higher rate. In this example. The optimal (C, γ) pair is $(2^{-3}, 2^{-7})$ that leads to a cross-validation rate superior than 0.8 (80%) as highlighted by the red mark.

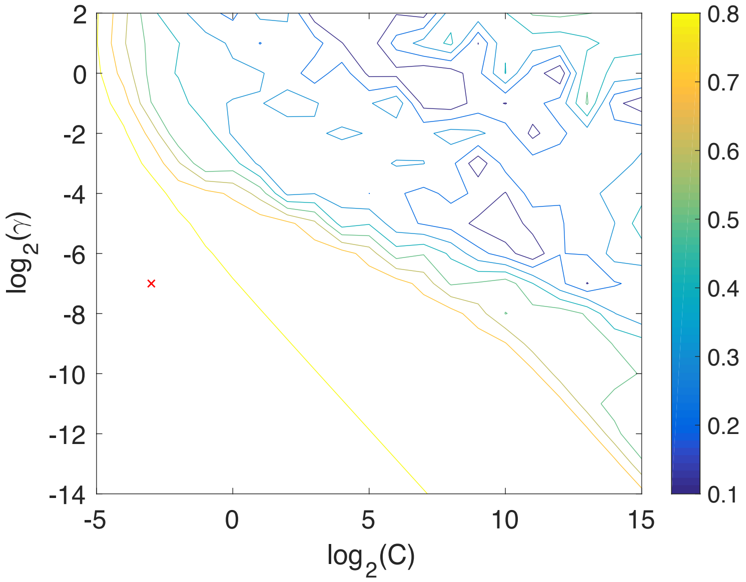


FIGURE 2.9 – Grid search on $C = 2^{(-5:15)}$ and $\gamma = 2^{(-14:2)}$ using the radial basis function (RBF) kernel and the features extracted from the LIRIS masking database objects.

2.5.2 Influence of the SVR kernel

The support vector regression provides the possibility to use four different kernels (linear, polynomial, RBF, and sigmoid) depending on the desired application. In our

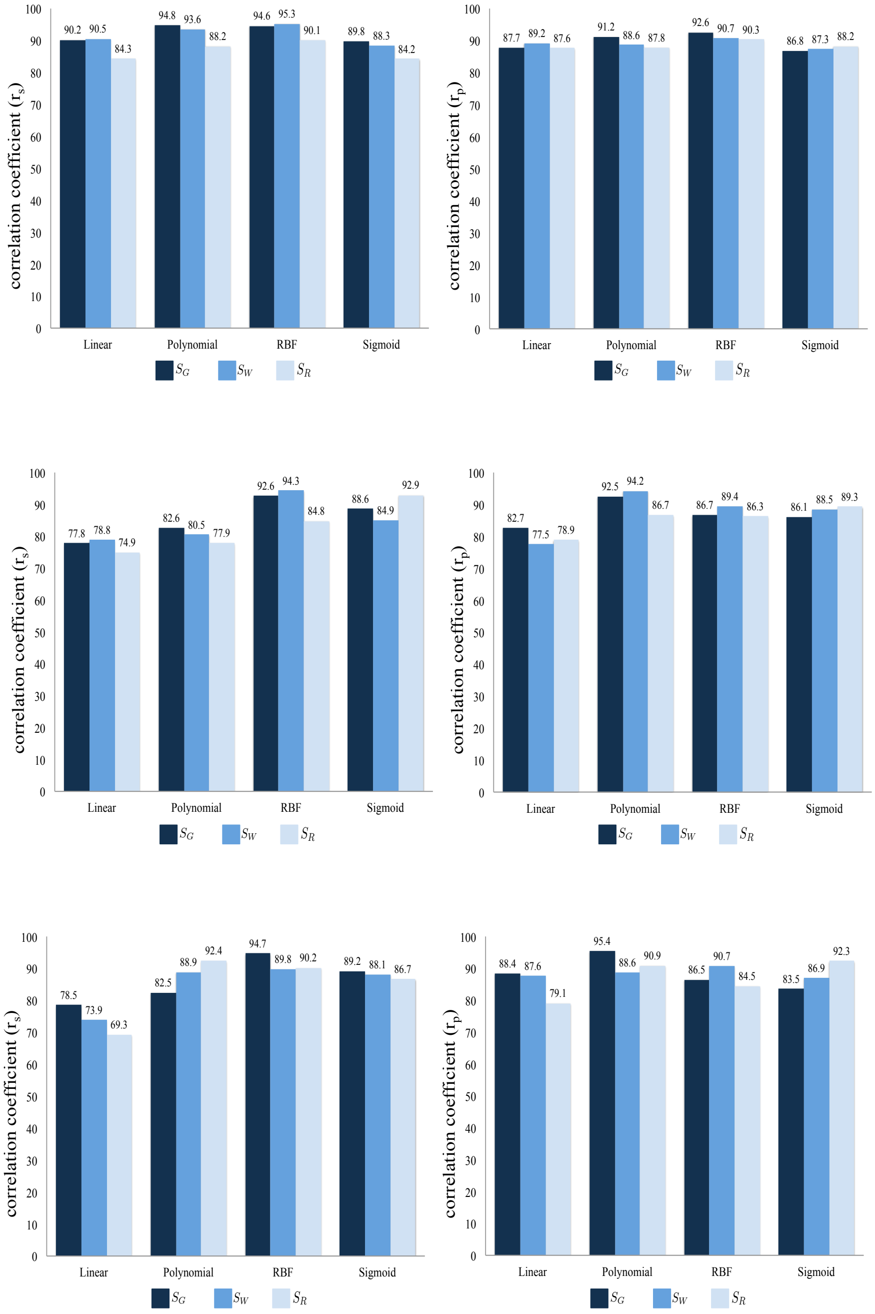


FIGURE 2.10 – Correlation coefficient r_s (%) and r_p (%) according to the tested kernels on the LIRIS masking database (first row), the general-purpose database (second row) and the UWB compression database (third row).

context, it is not evident beforehand which kernel to use to efficiently predict the objective quality score. The ultimate goal is to select the adequate kernel with selective parameters that conducts to the optimal SVR prediction on the three databases. Fig. 2.10 illustrates the Pearson and Spearman correlation coefficients estimated by using each kernel. Note that, these coefficients are computed from the intermediate quality scores S_G, S_W and S_R on the LIRIS masking database, the general-purpose database, and the UWB compression database .

The correlation values presented in Fig. 2.10 show that the quality scores vary from a kernel to another. Although there is not a huge difference between the predicted scores by the different kernels, we remark that the radial basis function (RBF) and the polynomial kernel are slightly efficient. On the other hand, the linear kernel generally provides lower scores compared to the other kernels. This justified by the fact that the RBF kernel can handle the non-linearity issue i.e. when the relation between the subjective scores and attributes is nonlinear. Whereas the linear kernel which is a special case of RBF can handle only the linear problems. In addition, the RBF kernel has fewer numerical difficulties and fewer hyper-parameters comparing to the polynomial kernel. Accordingly, the RBF seems to be the appropriate kernel for the SVR used in our method, as justified by its high correlation scores comparing to the other kernels.

2.5.3 The interest of the visual masking effect modulation

To demonstrate the interest of the visual masking modulation discussed in Section. 2.3.2, we compare the correlation coefficients obtained by our method with and without including the visual masking modulation. Table. 2.3 presents the correlation coefficients r_s (%) and r_p (%) of our method with and without the visual masking effect modulation on the three used databases.

Tableau 2.3 – Correlation coefficients r_s (%) and r_p (%) of our method with and without the visual masking effect modulation on LIRIS masking database, LIRIS/EPFL general-purpose database and the UWB compression database.

	Masking database		General-Purpose database		Compression database	
	Spearman r_s	Pearson r_p	Spearman r_s	Pearson r_p	Spearman r_s	Pearson r_p
Without visual masking	82.2	80.4	84.4	83.7	85.7	87.3
With visual masking	91.1	89.1	84.6	86.8	85.5	88.1

It is shown from Table. 2.3 that the visual masking modulation significantly improves the obtained correlations on the LIRIS masking database where the r_s and r_p coefficients increase by 8.9% and 8.7% respectively. On the General-Purpose database, the scores are slightly improved after the masking module. We notice also an improvement by 0.8% of the r_p coefficient on the UWB compression database, whereas the r_s coefficient slightly decreases by 0.2%. Therefore, the used visual masking module is very effective, especially on the LIRIS masking database which is manufactured to evaluate the visual masking effect modulation.

2.5.4 Comparison with the state of the art

In this section, we compare our MVQ assessment methods based on hand-crafted features and machine learning with several existing full reference and reduced reference methods. We call the first method NR-SVR (method based on dihedral angles and SVR), and the second method NR-GRNN (method based on curvature features and GRNN). The values of r_s and r_p from the compared objective MVQ methods on the four considered databases are listed in Table. 2.4, where the highest correlation coefficients have been highlighted in boldface. We note that the correlations on the whole corpus are computed for the whole database i.e. the statistical dependence is computed between the objective scores of all objects in the corpus and the corresponding MOS.

TABLEAU 2.4 – Correlation coefficients r_s (%) and r_p (%) of different objective methods on LIRIS masking database, LIRIS/EPFL general-purpose database and the UWB compression database.

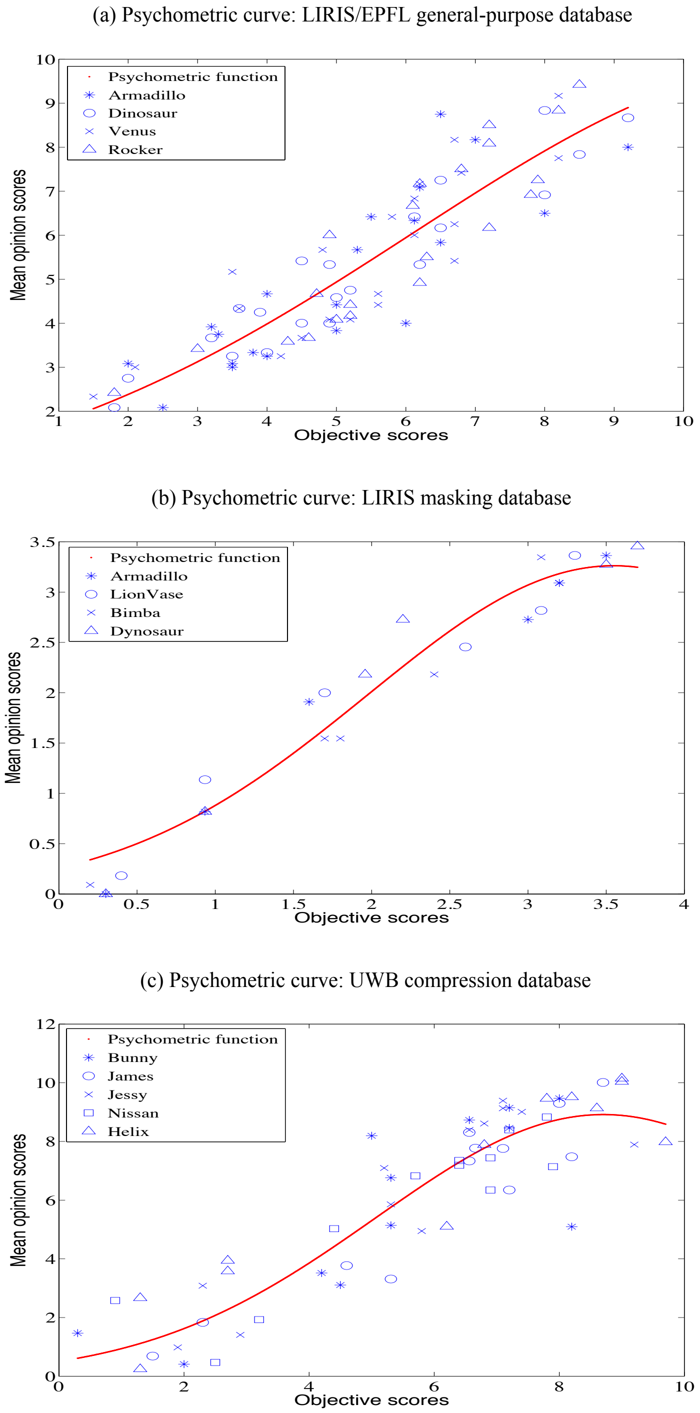
Method	Masking		General-purpose		Compression		Simplification	
	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Hausdorff Aspert *et al.* (2002)	26.6	20.2	13.8	11.4	24.5	14.0	49.4	25.5
RMS Cignoni *et al.* (1998)	48.8	41.2	26.8	28.1	52.0	49.0	64.3	34.4
MSDM Lavoué *et al.* (2006)	65.2	69.2	73.9	75.0	83.1	91.5	NA	NA
MSDM2 Lavoué (2011)	89.6	87.3	80.4	81.4	78.0	89.3	86.7	79.6
DAME Váša et Rus (2012)	68.1	58.6	76.6	75.2	85.6	**93.5**	NA	NA
GL1 Karni et Gotsman (2000)	42.0	39.6	33.1	35.5	66.9	70.6	NA	NA
GL2 Sorkine *et al.* (2003)	40.1	38.3	39.3	42.4	73.9	76.1	NA	NA
SEF Bian *et al.* (2009)	38.6	15.5	15.7	7.0	57.4	34.8	NA	NA
3DWPM1 Corsini *et al.* (2007)	29.4	31.9	69.3	61.8	81.9	84.1	NA	NA
3DWPM2 Corsini *et al.* (2007)	37.4	42.7	49.0	49.6	80.9	82.3	NA	NA
FMPD Wang *et al.* (2012)	80.2	80.8	81.9	83.5	81.8	88.8	87.2	**89.3**
TPDM Torkhani *et al.* (2014)	90.0	88.6	**89.6**	86.2	82.9	91.5	86.9	88.2
BMQI Nouri *et al.* (2017)	83.0	NA	78.1	NA	NA	NA	NA	NA
NR-SVR	**91.1**	**89.1**	84.6	86.8	85.5	88.1	**88.9**	87.6
NR-GRNN	90.2	82.4	86.2	**88.7**	**86.3**	86.7	87.7	88.0

It is remarkable that the classical metrics based on geometric distance HD and RMS generally fail to reflect the perceived visual quality, and do not correlate well with human perception. This failure is because these metrics compute a pure geometric distance neglecting the main operations of the HVS.
The proposed no-reference method NR-SVR has the highest Spearman and Pearson correlation values on the LIRIS masking database (whole corpus $r_s = 91.1$ and $r_p = 89.1$) and overcomes the most effective methods (TPDM, FMPD, and MSDM2). This database is manufactured to evaluate the visual masking effect and test the capacity of an objective MVQ method to capture it. The high results and performances provided by the proposed method confirm that the visual masking effect is well captured and the modulation used in this work is very effective. NR-GRNN has also good correlation scores with $r_s = 90.2$ and $r_p = 82.4$.
On the general-purpose database, NR-SVR method provide high correlation scores that contend the most effective methods i.e high Pearson correlation score ($r_p = 86.8$) and the second-highest Spearman correlation score ($r_s = 84.6$) behind TPDM method ($r_s = 89.6$). NR-GRNN provides the highest Pearson correlation score ($r_p = 88.7$) on this database. Comparing to the other databases, the general-purpose database contains the highest number of distorted models (21 distorted versions for each model as well as a variety of distortion types). The high correlation scores provided by the proposed methods in this database appear to be a good indicator for their forcefulness in MVQ assessment.
On the UWB compression database, NR-SVR has competitive scores and it outperforms several effective methods FMPD, 3DWPM1, 3DWPM1, TPDM, and MSDM2 in terms of the overall Pearson and Spearman coefficients. In particular, the overall r_s of our method is better than TPDM method (85.5% for our method vs. 82.9% for TPDM). Even if the obtained overall r_p is significant, several methods outperform this score for such a database. This is because the obtained objective scores of the five models in this database are not exactly in the same rank comparing to other effective methods. on the other hand, NR-GRNN provides the highest Spearman coefficient on this database $r_s = 86.3\%$.
The good performance of NR-SVR method can be also illustrated by the scatter plots of MOSs versus the predicted scores in Fig 2.11. It is shown that the fitted psychometric function has good generalization capability on the three databases, where the generated "objective scores vs MOS" points are very close to the psychometric curve.

2.5.5 Execution time

For real-time applications the execution time is an important factor. To demonstrate the eligibility of our method for real-time applications, we present in Table 2.5 the computational time regarding the tests conducted on the three databases. Note that the execution time presented in this table is obtained after testing all the objects for each database. The runtime environment is MATLAB R2016a on 1,8 GHz CPU Intel Core i5 macOS Sierra.

According to the above evaluation and comparison, we can see that the proposed method is quite robust and effective in terms of predicting the objective quality score of a distorted mesh, as reflected by the high and competitive scores comparing to existing



(a) Psychometric curve: LIRIS/EPFL general-purpose database

(b) Psychometric curve: LIRIS masking database

(c) Psychometric curve: UWB compression database

FIGURE 2.11 – Scatter plots of the mean opinion scores (MOS) versus the objective scores obtained from the proposed method. (a) LIRIS/EPFL general-purpose database. (b) LIRIS masking database. (c) the UWB compression database.

Tableau 2.5 – Execution time.

Database	LIRIS Masking	General-Purpose	UWB Compression
Execution time (s)	123	566	182

well-known methods. In addition, we recall that our proposed method is a no-reference MVQ assessment method, and do not require reference meshes for the quality estimation. Accordingly, it can be appropriate for practical situations and real-time applications. Furthermore, the execution time can also justify the eligibility of our method for such applications.

2.6 Conclusion

In most practical situations, there is limited access to the information related to the reference and the distortion type. For these reasons, the development of a no-reference mesh visual quality (MVQ) approach is a critical issue, and more emphasis needs to be devoted to blind methods. Several full reference and reduced reference methods have been proposed to estimate the perceived visual quality of 3D meshes. However, blind mesh quality assessment is not well addressed in the literature. In this chapter, we presented two metrics to blindly assess the perceived visual quality of distorted meshes. The main goal of these contributions is to consider the perceptual information of the 3D mesh by extracting hand-crafted features. Besides, we use machine learning techniques for feature learning and quality prediction.

The first method that we called NR-SVR extract dihedral angles as relevant information that describes the structural information. The extracted feature vector is then modulated with a visual masking function which is an important characteristic of the human visual system. The obtained vector is then modeled by three statistical distributions : Gamma, Weibull, and Rayleigh to construct three sets of feature vectors with only statistical parameters instead of using whole features values. This step is very crucial to reduce computational time. Once the feature vectors are constructed, the proposed scheme predicts three intermediate scores using the support vector regression (SVR) according to the three distributions. Finally, the overall objective quality score is computed by a weighted sum of the intermediate scores. The second method that we called NR-GRNN based on the general regression neural network (GRNN) fed by the mean curvature, which considered as a perceptual relevant feature representing the visual aspect of a 3D mesh.

The obtained results are promising and show the effectiveness of machine learning for the MVQ assessment task as proven by the experimental results. Recently, deep learning has attracted the attention of many researchers. They have been successfully employed in various computer vision applications allowing them to reach high performances. In this context, we are going further with machine learning techniques and we take advantage of deep learning to develop an MVQ assessment metric for better estimation. The next chapter presents an introduction to convolutional neural networks and their application to MVQ. In the next chapter, we introduce the use of deep learning for better estimation.

Chapitre

3 Convolutional neural network for mesh visual quality assessment

Sommaire

3.1 Introduction

In the last chapter, we introduced machine learning to blindly assess the perceived quality of distorted meshes. The classical methods SVR and neural networks provide promising results and demonstrate the power of these tools in our application. This leads us to go further and use more sophisticated networks to improve the quality estimation.
Convolutional neural networks (CNN) have recently attracted the attention of many researchers . They have been succesfully employed in various computer vision applications allowing to reach high performances LeCun *et al.* (2015). One of their main advantage over classical neural networks is that they adequately considers the spatial structure of

the input data. Moreover, CNN allows the important property of weights sharing between the convolutional layers which restrict the number of parameters to learn. Their use in blind image quality assessment (BIQA) has shown notable improvment in terms of the correlation with the human judgment Zhang *et al.* (2016). However, it has not yet been exploited for MVQ assessment. Indeed, studies in mesh quality assessment tend more to adopt FR and RR approaches, since they usually perform better than blind methods.
In this chapter, we propose a method based on deep learning for the mesh visual quality assessment without reference. For a given 3D modl, we render 2D views from the 3D mesh. After that, the views are split into small patches of a fixed size. Using the prepared data, the first method to estimate to perceived visual quality is to use a CNN architecture from scratch. The second method is to use pre-trained CNNs, Each network is fine-tuned and produces a feature vector. The Compact Multi-linear Pooling (CMP) is used afterward to fuse the retrieved vectors into a global feature representation. Finally, fully connected layers followed by a regression module are used to estimate the quality score.

In the first section of this chapter, we introduce deep learning, especially, convolutional neural network and its applications. In the second section, we describe how we prepare the input data used in our approach. The third section is dedicated to a detailed description of the first CNN-based method using a network from scratch. Then, in the fourth section, we present a developed method that uses fine-tuned networks and compact multilinear pooling for quality estimation. Extensive experiments are executed and discussed in the last section of this chapter.

3.2 Convolutional neural networks

Before describing the different elements of Convolutional neural networks, it is important to present some definitions and background.

3.2.1 Background

Artificial Neural Networks (ANNs) Hassoun *et al.* (1995) are computational processing systems of which are heavily inspired by the way biological nervous systems (such as the human brain) operate. ANNs are mainly comprised of a high number of interconnected computational nodes (referred to as neurons), of which work entwine in a distributed fashion to collectively learn from the input in order to optimize its final output. The input is usually loaded in the form of a multidimensional vector to the input layer of which will distribute it to the hidden layers. The hidden layers will then make decisions from the previous layer and weigh up how a stochastic change within itself detriments or improves the final output, and this is referred to as the process of learning. Having multiple hidden layers stacked upon each-other is commonly called deep learning. The two key learning paradigms in image processing tasks are supervised and unsupervised learning.
Supervised learning Caruana et Niculescu-Mizil (2006) is learning through pre-labelled inputs, which act as targets. For each training example, there will be a set of input values (vectors) and one or more associated designated output values. The goal of this form of

training is to reduce the models' overall classification error, through a correct calculation of the output value of training example by training.

Unsupervised learning Barlow (1989) differs in that the training set does not include any labels. Success is usually determined by whether the network is able to reduce or increase an associated cost function. However, it is important to note that most image-focused pattern recognition tasks usually depend on classification using supervised learning.

Overfitting Sarle (1996) is basically when a network is unable to learn effectively due to a number of reasons. It is an important concept of most, if not all machine learning algorithms and it is important that every precaution is taken to reduce its effects. If the models were to exhibit signs of overfitting then we may see a reduced ability to pinpoint generalized features for not only our training dataset but also our test and prediction sets. This is the main reason behind reducing the complexity of ANNs. The fewer parameters required to train, the less likely the network will overfit and improve the predictive performance of the model.

Deep learning is a set of learning methods attempting to model data with complex architectures combining different non-linear transformations. The elementary bricks of deep learning are the neural networks, that are combined to form the deep neural networks. These techniques have enabled significant progress in the fields of sound and image processing, including facial recognition, speech recognition, computer vision, automated language processing, text classification. There exist several types of architectures for neural networks :

- The multilayer perceptrons, that are the oldest and simplest ones
- The Convolutional Neural Networks (CNN), particularly adapted for image processing
- The recurrent neural networks, used for sequential data such as text or times series.

They are based on a deep cascade of layers. They need clever stochastic optimization algorithms, and initialization, and also a clever choice of the structure. They lead to very impressive results, although very few theoretical foundations are available till now.

Convolutional Neural Network has had groundbreaking results over the past decade in a variety of fields related to pattern recognition ; from image processing to voice recognition. The most beneficial aspect of CNNs is reducing the number of parameters in ANN. This achievement has prompted both researchers and developers to approach larger models to solve complex tasks, which was not possible with classic ANNs. The most important assumption about problems that are solved by CNN should not have spatially dependent features. In other words, for example, in a face detection application, we do not need to pay attention to where the faces are located in the images. The only concern is to detect them regardless of their position in the given images. Another important aspect of CNN is to obtain abstract features when input propagates toward the deeper layers. For example, in image classification, the edge might be detected in the first layers, and then the simpler shapes in the second layers, and then the higher-level features. From the input raw image vectors to the final output of the class score, the entire network will still express a single perceptive score function (the weight). The last layer will contain loss functions associated with the classes, and all of the regular tips and tricks developed for traditional ANNs still apply. The only notable difference between CNNs and traditional ANNs is that CNNs are primarily used in the field of pattern recognition within images.

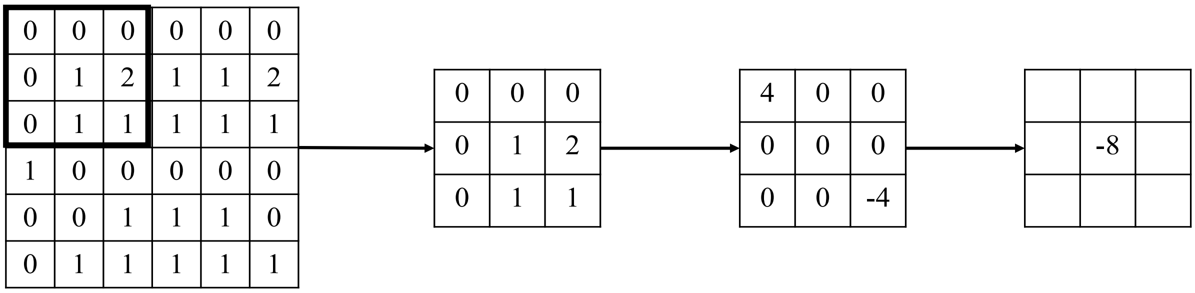
This allows us to encode image-specific features into the architecture, making the network more suited for image-focused tasks - whilst further reducing the parameters required to set up the model.

3.2.2 Convolutional neural network elements

A Convolutional Neural Network is composed of several kinds of layers, that are described in this section : convolutional layers, pooling layers, and fully connected layers. Besides, we present important parameters in a CNN architecture i.e stride, padding, the notion of non-linearity and CNN features.

Convolutional layer

As the name implies, the convolutional layer plays a vital role in how CNNs operate. The parameters of layers focus on the use of learnable kernels. These kernels are usually small in spatial dimensionality but spread along the entirety of the depth of the input. When the data hits a convolutional layer, the layer convolves each filter across the spatial dimensionality of the input to produce a 2D activation map Papandreou *et al.* (2015); Krizhevsky *et al.* (2012). As we glide through the input, the scalar product is calculated for each value in that kerne (Fig. 3.1). From this, the network will learn kernels that 'fire' when they see a specific feature at a given spatial position of the input. These are commonly known as activations. Every kernel will have a corresponding activation map,



0	0	0	0	0	0
0	1	2	1	1	2
0	1	1	1	1	1
1	0	0	0	0	0
0	0	1	1	1	0
0	1	1	1	1	1

0	0	0
0	1	2
0	1	1

4	0	0
0	0	0
0	0	-4

	-8	

FIGURE 3.1 – A visual representation of a convolutional layer. The center element of the kernel is placed over the input vector, of which is then calculated and replaced with a weighted sum of itself and any nearby pixels.

which will be stacked along the depth dimension to form the full output volume from the convolutional layer. Parameter sharing works on the assumption that if one region feature is useful to compute at a set spatial region, then it is likely to be useful in another region. If we constrain each individual activation map within the output volume to the same weights and bias, then we will see a massive reduction in the number of parameters being produced by the convolutional layer. As a result of this as the backpropagation stage occurs, each neuron in the output will represent the overall gradient of which can be totaled across the depth. Thus only updating a single set of weights, as opposed to every single one.

Stride

In fact, CNN has more options which provide a lot of opportunities to even decrease the parameters more and more, and at the same time reduce some of the side effects. One of these options is stride. In the above-mentioned example, it is simply assumed that the next layer's node has lots of overlaps with its neighbors by looking at the regions. We can manipulate the overlap by controlling the stride. Fig. 3.2, shows a given 7 × 7 image. If we move the filter one node every time, we can have a 5 × 5 output only. Note that the output of the three left matrices has an overlap (and three middle ones together and three right ones also). However, if we move and make every stride 2, then the output will be 3 × 3. Not only overlap but also the size of the output will be reduced Wu (2017). Eq. 3.1, formalize this, given the image $N \times N$ dimension and the filter size of the $F \times F$, and the stride s, the output size O.

$$O = 1 + \frac{N - F}{s} \tag{3.1}$$

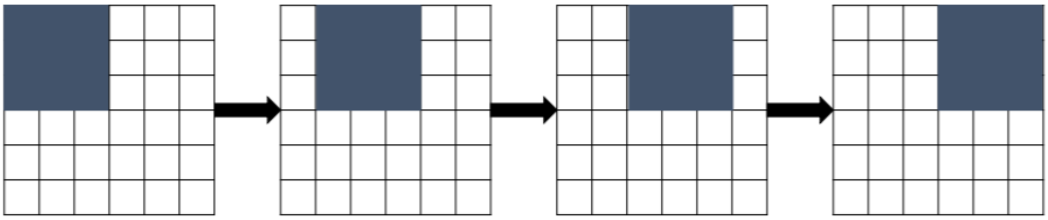


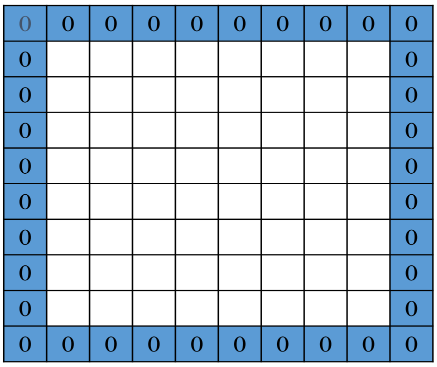
Figure 3.2 – Stride 1, the filter window moves only one time for each connection

Padding

One of the drawbacks of the convolution step is the loss of information that might exist on the border of the image. Because they are only captured when the filter slides, they never have the chance to be seen. A very simple, yet efficient method to resolve the issue, is to use zero-padding. The other benefit of zero padding is to manage the output size. For example, in Fig. 3.2, with $N = 7$ and $F = 3$ and stride 1, the output will be 5 × 5 (which shrinks from a 7 × 7 input). Zero-padding is the simple process of padding the border of the input and is an effective method to give further control as to the dimensionality of the output volumes. It is important to understand that through using these techniques, we will alter the spatial dimensionality of the output of the convolutional layer. To calculate this, we can make use of the following equation :

$$O = 1 + \frac{N + 2P - F}{s} \tag{3.2}$$

Where P is the number of the layers of the zero padding (e.g. $P = 1$ in Fig. 3.3), This padding idea helps us to prevent network output size from shrinking with depth. Therefore, it is possible to have any number of deep convolutional networks Wu (2017).

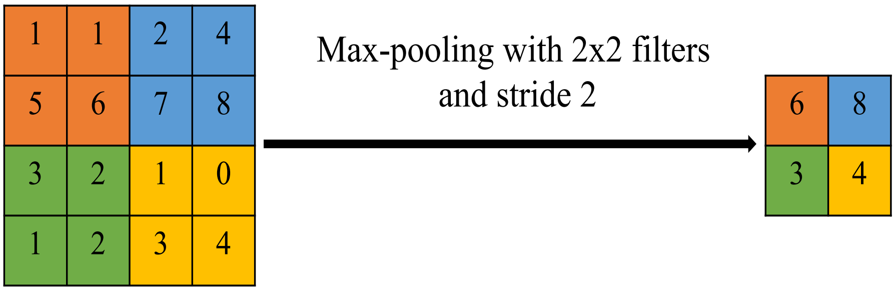


0	0	0	0	0	0	0	0	0	0
0									0
0									0
0									0
0									0
0									0
0									0
0									0
0									0
0	0	0	0	0	0	0	0	0	0

Figure 3.3 – Example of zero padding with $P = 1$.

Pooling layer

CNN also has pooling layers, which allow reducing the dimension, also referred to as subsampling, by taking the mean or the maximum on patches of the image (mean pooling or max-pooling). Like the convolutional layers, pooling layers acts on small patches of the image, we also have a stride. In the image processing domain, it can be considered as similar to reducing the resolution. Pooling does not affect the number of filters. Max-pooling is one of the most common types of pooling methods. It partitions the image to sub-region rectangles, and it only returns the maximum value of the inside of that sub-region. One of the most common sizes used in max-pooling is 2×2. As can see in Fig. 5.2, when pooling is performed in the top-left 2×2 blocks (pink area), it moves 2 and focuses on the top-right part. This means that stride 2 is used in pooling. To avoid down-sampling, stride 1 can be used, which is not common. It should be considered that down-sampling does not preserve the position of the information. Therefore, it should be applied only when the presence of information is important (rather than spatial information). Moreover, pooling can be used with non-equal filters and strides to improve the efficiency. For example, a 3×3 max-pooling with stride 2 keeps some overlaps between the areas. Zeiler et Fergus (2014); Giusti *et al.* (2013).



1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Figure 3.4 – The max-pooling with 2×2 filter and stride 2 lead to down-sampling of each 2×2 blocks is mapped to 1 block (pixel).

Fully-connected layer

After several convolution and pooling layers, the CNN generally ends with several fully connected layers. The tensor that we have at the output of these layers is transformed into a vector and then we add several perceptron layers. The fully-connected layer contains neurons of which are directly connected to the neurons in the two adjacent layers, without being connected to any layers within them. This is analogous to the way that neurons are arranged in traditional forms of ANN. The major drawback of a fully-connected layer is that it includes a lot of parameters that need complex computational in training examples. Therefore, we try to eliminate the number of nodes and connections. The removed nodes and connection can be satisfied by using the dropout technique. For example, LeNet and AlexNet designed a deep and wide network while keeping the computational complex constant Russakovsky *et al.* (2015).

Overall CNN architecture

After the different types of layers composing a CNN are described. We present here how these layers are combined to form the architecture of the network. Choosing an architecture is very complex and this is more engineering than an exact science. It is therefore important to study the architectures that have proved to be effective and to draw inspiration from these famous examples. In the most classical CNN, we chain several times a convolution layer followed by a pooling layer and we add at the end fully connected layers. The LeNet network, proposed by the inventor of the CNN, Yann LeCun LeCun *et al.* (1998) is of this type, as shown in Fig. 3.5. This network was devoted to digit recognition. It is composed only on few layers and few filters, due to the computer limitations at that time.

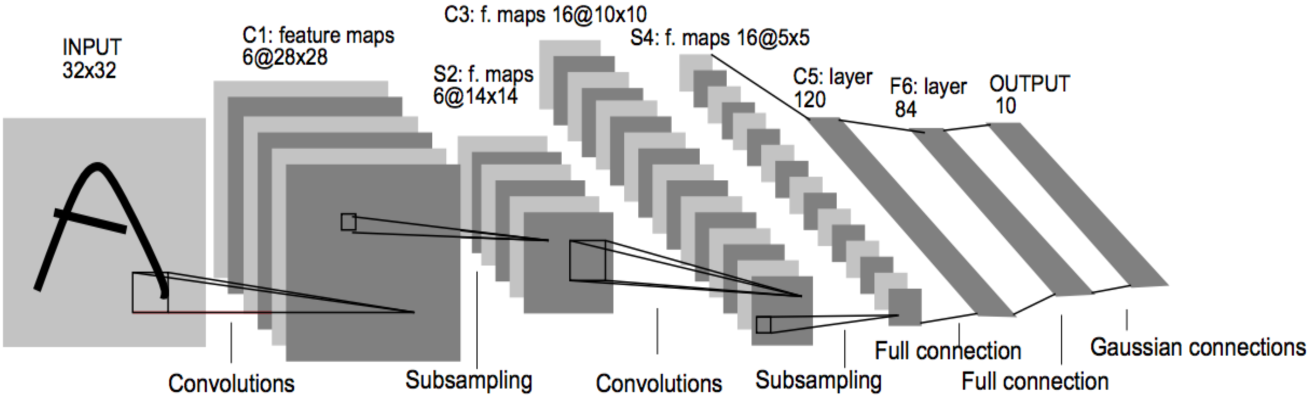


FIGURE 3.5 – Architecture of the network Le Net LeCun *et al.* (1998).

A few years later, with the appearance of GPU (Graphical Processor Unit) cards, much more complex architectures for CNN have been proposed, like the network AlexNet Krizhevsky *et al.* (2012) that won the ImageNet competition and for which a simplified version is presented in Fig. 3.6. This competition was devoted to the classification of one million color images onto 1000 classes. The resolution of images was 224 × 224. AlexNet is composed of 5 convolution layers, 3 max-pooling 2×2 layers, and fully connected layers. As showed in Fig. 3.6, the kernel shape of the first convolution layer is $(11, 11, 3, 96)$ with a stride of $s = 4$, and the first output shape is $(55, 55, 96)$.

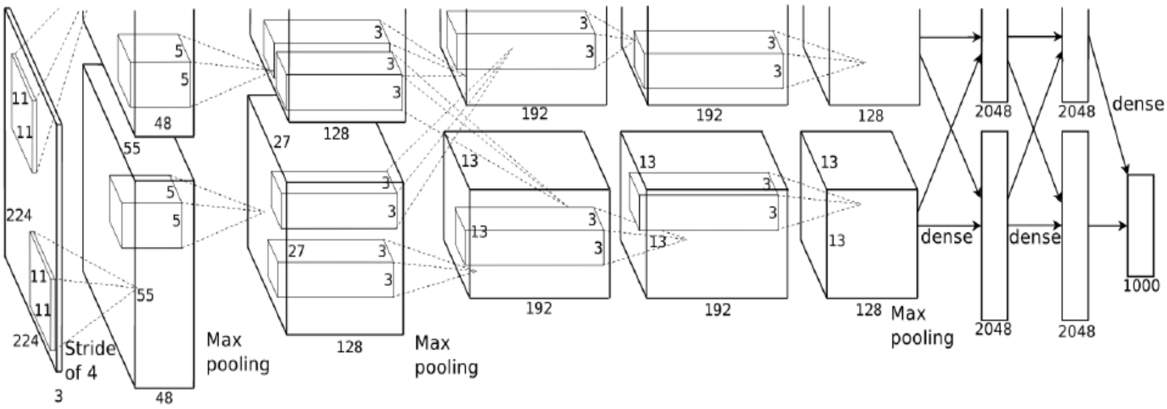


FIGURE 3.6 – Architecture of the network AlexNet Krizhevsky *et al.* (2012).

In the next sections of this chapter, we describe in detail the proposed schemes using CNN architectures to objectively estimate the perceived visual quality of 3D meshes. We begin by presenting how the input data are prepared and then the learning step and the quality estimation.

3.3 Learning data preparation

3.3.1 Mesh views rendering

The first step of the proposed scheme consists of rendering 2D projections. This strategy permits to represent the 3D mesh from multiple views. To do so, virtual cameras are fixed at different angles around the 3D mesh according to the axes X and Y. As illustrated in Fig. 3.7, the centroid of the 3D object is placed at the origin of the coordinate system. The coordinates (x_i, y_i) of the virtual cameras are obtained by varying the angles $x \in [0, 2\pi]$ and $y \in [0, 2\pi]$ by $\frac{\pi}{6}$ (30 degrees). Twelve angles are obtained for each axis. Hence, for each combination of x_i and y_i a virtual camera is placed and a 2D projection is obtained. In total, 144 projections are obtained from each 3D mesh. We note that we use only the axes X and Y since using also the Z axis duplicates the views and provides redundant and useless information. In our projection strategy, the obtained projections describe the 3D object from all-important views. Moreover, in view-based 3D shape retrieval also only X- and Y-rotations are used to generate 2D views of 3D objects Bai *et al.* (2017); Zhu *et al.* (2017).

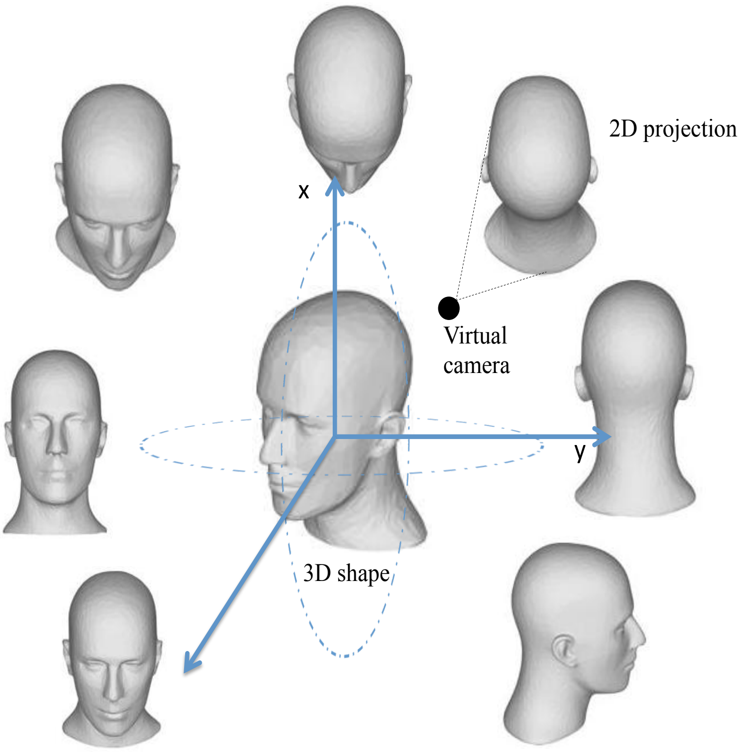


FIGURE 3.7 – Mesh views rendering, a virtual camera is placed for each combination of x and y. 144 projections are obtained for each 3D mesh.

3.3.2 Input patches normalisation

The next step is to apply a simple local contrast normalization on the selected patches. The normalized value $\hat{I}(i,j)$ of a pixel $I(i,j)$ at location (i,j) is computed as follows :

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + c} \tag{3.3}$$

$$\mu(i,j) = \frac{1}{(2M+1)\times(2N+1)} \sum_{m=-M}^{m=M} \sum_{n=-N}^{n=N} I(i+m, j+n) \tag{3.4}$$

$$\sigma(i,j) = \sqrt{\sum_{m=-M}^{m=M} \sum_{n=-N}^{n=N} (I(i+m,j+n) - \mu(m,n))^2} \tag{3.5}$$

where c is a constant that prevents instabilities from dividing by zero. M and N are the normalization window sizes. The used normalization is crucial to make the trained networks robust to illumination and contrast variation by decreasing the effect of the saturation problem Mittal *et al.* (2012).

3.4 CNN from scratch for MVQ assessment

Using the prepared data, the first method to estimate the perceived visual quality is to use a CNN network from scratch.
The CNN is fed by the normalized patches of size 32 × 32. Fig. 3.8 shows the different layers of the used CNN.

Note that several network configurations have been tested to choose the best architecture for our method. The elaborated architecture in this section is the one that led to the best results (more details in Section 3.6).

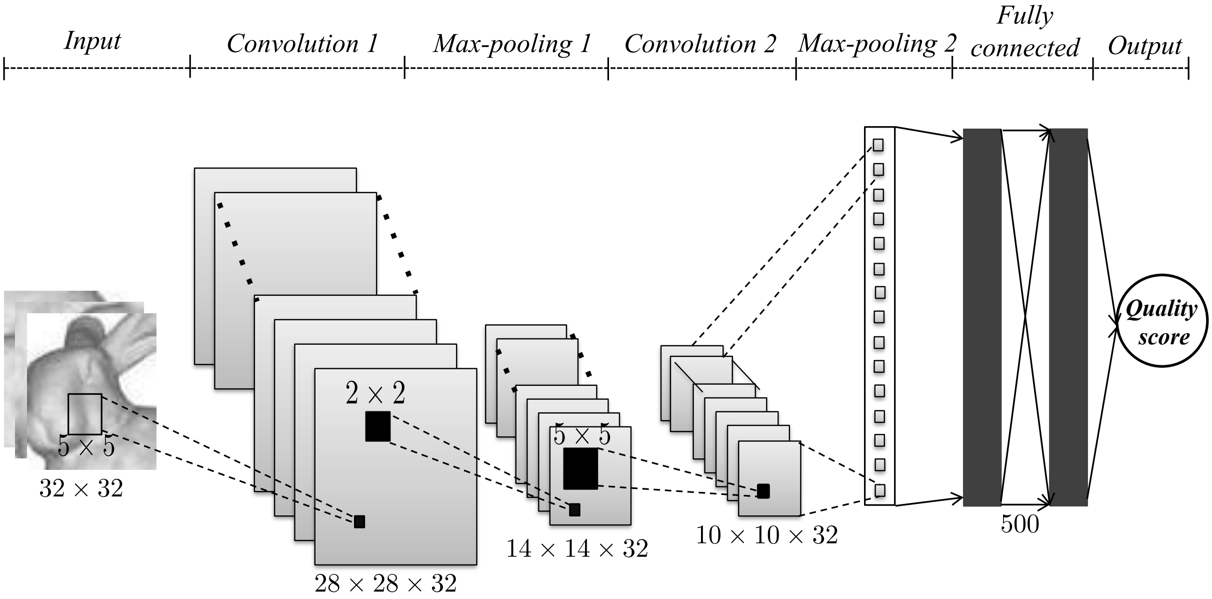


FIGURE 3.8 – Convolutional neural network configuration of the proposed method.

The first layer of the network is a convolutional layer, it filters the input patch with 32 kernels of size (5 × 5). The convolution process is defined as follows :

$$Y_i = W_i * X + b_i, \quad i = 1, 2, ..., N \tag{3.6}$$

where X is the input patch of the CNN and $*$ is the convolution operation. $\{W_i\}_{i=1}^{N}$ denotes the convolutional kernels and $\{b_i\}_{i=1}^{N}$ are the biases values. 32 feature maps (28 × 28) are generated by this layer.

The second layer in our network is a max-pooling layer. It applies the max pooling operation on the feature maps generated by the previous layer to reduce the dimension of the filter responses. The max-pooling operation is defined as follows :

$$M_{x,y}^{n} = \max_{(x,y)\in\Omega} (Y_{x,y}^{n}) \tag{3.7}$$

where $M_{x,y}^{n}$ denotes the output of the max-pooling layer (maximum values). $n = 1, 2, ..., N$ where N is the number of filters. Ω is a local window used in the pooling operation. In this layer, we use a local window of size 2 × 2 with a stride equals to 2. This provides 32 feature maps (14×14).

Another convolutional layer is introduced with 32 kernels (5×5). This layer provides 32 maps of size 10×10. It filters the feature maps obtained from the previous layer using the same process in Eq.3.6.

After the second convolutional layer, we introduce another max-pooling layer of size 10×10. The size of the local window, in this case, is the same as the maps produced by the previous layer, in other words, each feature map will be reduced to a single value and its output is thus a vector of size 1×32.

To estimate the perceived quality scores of a given distorted 3D model, the obtained feature vectors are then used to feed two fully connected layers with 500 neurons each. In our work, we adopt the non-linear activation function ReLU (Rectified Linear Units) Nair et Hinton (2010).

3.5 Fine-tuned networks and compact multilinear pooling for MVQ

The second proposed method is to fuse automatically learned features extracted from CNN models. To this end, we fine-tune three well-known pre-trained CNN models (AlexNet, VGG, and ResNet). A feature vector is then extracted for each network and a defined combination is applied to obtain a global feature vector to be used for the quality prediction. In the next, we give more details about the used networks and the combination strategies.

3.5.1 Fine-tuned networks

The CNN models used in this work are briefly described as follows :

— **AlexNet Krizhevsky (2014) :** This CNN model, proposed by Alex Krizhevsky, is the winner of the Image Large Scale Visual Recognition Challenge (ILSVRC) in 2012. It consists of five convolutional layers, max-pooling layers, and three fully-connected layers. The dropout regularization method is used in the fully connected layers to prevent overfitting. Also, the authors highlight the use of the ReLU function and the overlap in the max-pooling layers.

— **VGG Simonyan et Zisserman (2014) :** VGG network is a deep CNN proposed by the Oxford Visual Geometry Group. The network achieved successful performance in ILSVRC 2014. Several versions of VGG have been developed with different convolutional layers : VGG11, VGG13, VGG16, and VGG19. In our method, we use VGG16. This network consists of 13 convolutional layers with max-pooling and three fully connected layers.

— **ResNet He *et al.* (2016) :** The residual Neural Network (ResNet) has been proposed by Kaimimg He et al in 2015. The network is developed to make the training of deep networks easier by accelerating the speed of the training. Different versions of ResNet have been proposed : ResNet 18, ResNet 34, ResNet 52, and others. The used network consists of 16 convolutional layers with max-pooling and two fully connected layers.

It is worth noting that the last layer is a regression since the quality scores are seen as "continuous values". In addition, the input of the pre-trained network is adjusted to be fed by patches of a fixed size (patch-size = 32×32).

3.5.2 Compact multilinear pooling

Once the feature vectors are extracted from the above-described models, we combine them using the concept of Compact Bi-linear Pooling (CBP) Fukui *et al.* (2016) that computes the outer product of two feature vectors u and v. The authors demonstrate that this outer product can be seen as a convolution ($\circledast$) when the Count Sketch projection function is applied. This latter aims to project the feature vectors into a lower-dimensional feature space. Moreover, as the convolution of two vectors is equivalent to the element-wise product in the frequency domain, the outer product $u \circledast v$ can be finally rewritten as $FFT^{-1}(FFT(u) \odot FFT(v))$, where $\odot$ refers to the element-wise product and FFT designs the Fast Fourier Transform. This combination is very important because it permits the interaction of all elements of the vectors in a multiplicative way without a high computational cost.

To combine more than two vectors, the process is extended to Compact Multi-linear Pooling (CMP) Algashaam *et al.* (2017) as depicted in Fig. 3.9.a. The CMP combination consists firstly of projecting the feature vectors to a lower-dimensional feature space through a Count Sketch projection function. After computing the FFT of each considered feature vector, we multiply the obtained spectra and apply the Inverse Fast Fourier Transform (IFFT) to obtain a single feature vector.

In Section 3.6.5), the performance of CMP strategy is compared to some common combination, described as follows :

- **Concatenation (see Fig. 3.9.b) :** The simplest way to combine vectors is to concatenate them. It allows all the elements to interact in the learning, however, the result vector contains more elements and can slow down the prediction time.
- **Element-wise multiplication (see Fig. 3.9.c) :** The feature vectors can be combined by multiplying their elements, the result vector is of a smaller size than the concatenated one, however, not all the elements interact together in the learning process and some information can be lost.
- 1×1 **convolution :** another type of combination is to use 1×1 convolution filter. It creates a linear projection of the features and is used to reduce the data size.

CMP is interesting because it allows the interaction of all elements of the vectors in a multiplicative way, which is not the case in the element-wise multiplication. Besides, unlike the bilinear pooling, the CMP projects the outer product to a lower-dimensional space that leads to lower computation time.

After the CMP strategy is applied, a global feature vector is obtained, and it is fed to fully connected layers followed by a regression layer for the quality score estimation.

3.6 Experimental results

The CNN involves several parameters and provides an important degree of freedom to design an effective architecture for a specific application. To retrieve the best network architecture for our application, several parameters have been tested to investigate how the performance is affected and choose the best configuration. To do so, we first fix the patch

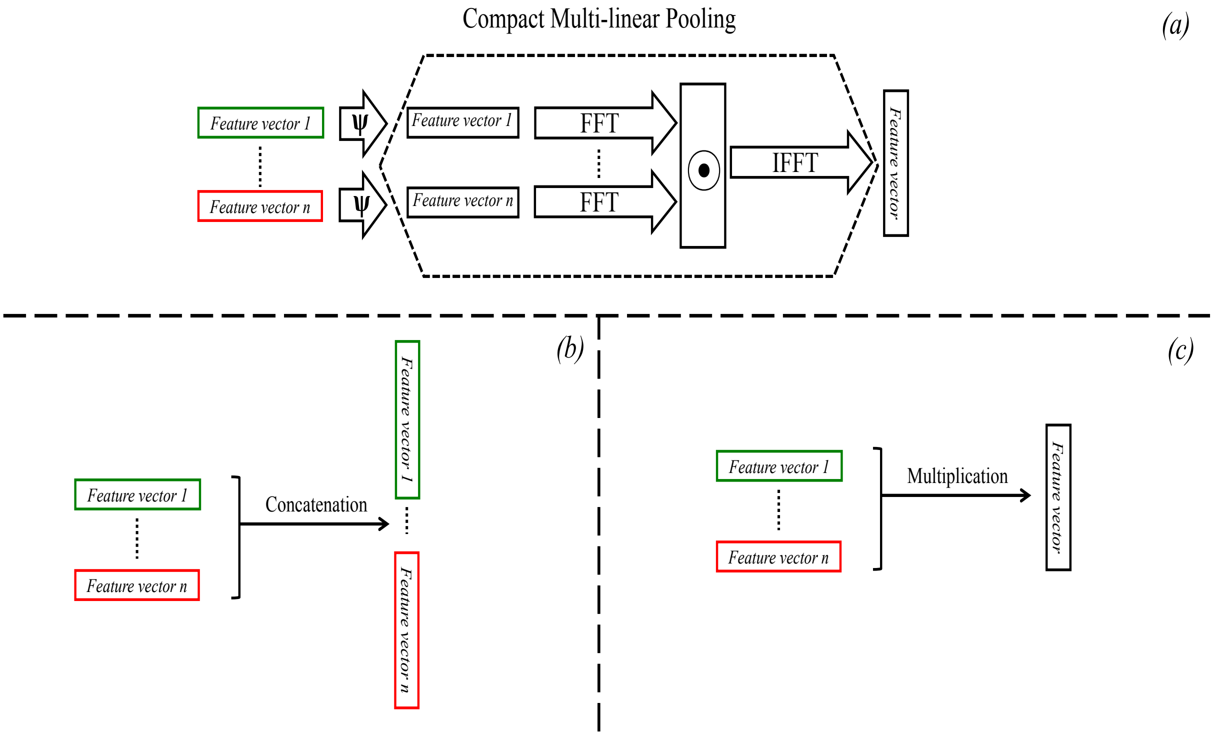


FIGURE 3.9 – Combination strategies.

size (32×32) and the kernel size (5×5) while testing the network with a different number of convolutional kernels. After that, we adjust the kernel size while fixing the number of kernels and the patch size. Finally, we examine the performance of the network by varying the patch size while fixing the number and the size of kernels to the best configuration obtained. We also compare the performance obtained by different combination strategies of the fine-tuned networks, and we present the experimental results and comparative analysis on mesh visual quality assessment state of the art.

3.6.1 Effect of the number of filters in the convolutional layers

As the convolutional layer is the core building of a CNN. The number of convolutional adopted for this layer could also influence the performance of the network. To demonstrate the influence of this parameter, we test the ability of our network in predicting the visual quality by using a variety of convolution kernels while fixing the other parameters. Table. 4.2 presents the performance of the network regarding the correlation coefficients with respect to the size of convolution kernels.

It is shown from Table. 4.2 that the number of kernels significantly affects the performance of the network. Using 32 kernels instead of 10 leads to an important improvement, however, using more kernels than 32 decreases the performance of the network in predicting the visual quality.

Tableau 3.1 – Performance of the network with respect to the number of kernels.

Network configuration	Input : (32 × 32) conv1-10 (5 × 5) max-pool1 (2 × 2) conv2-10 (5 × 5) max-pool2 (10 × 10) FC-500	Input : (32 × 32) conv1-32 (5 × 5) max-pool1 (2 × 2) conv2-32 (5 × 5) max-pool2 (10 × 10) FC-500	Input : (32 × 32) conv1-50 (5 × 5) max-pool1 (2 × 2) conv2-50 (5 × 5) max-pool2 (10 × 10) FC-500
r_s	88.6	**93.3**	92.4
r_p	89.7	**92.2**	91.2

3.6.2 Effect of the size of filters

Another parameter tested in our experiments is the size of the convolution kernels. To do so, we fix the input patch size and the number of convolution kernels while testing different sizes of the kernels. Table. 4.3 presents the performance of the network regarding the correlation coefficients with respect to the number of convolution kernels.

Tableau 3.2 – Performance of the network with respect to the size of convolutional kernels.

Network configuration	Input : (32 × 32) conv1-32 (3 × 3) max-pool1 (2 × 2) conv2-32 (3 × 3) max-pool2 (13 × 13) FC-500	Input : (32 × 32) conv1-32 (5 × 5) max-pool1 (2 × 2) conv2-32 (5 × 5) max-pool2 (10 × 10) FC-500	Input : (32 × 32) conv1-32 (7 × 7) max-pool1 (2 × 2) conv2-32 (7 × 7) max-pool2 (7 × 7) FC-500	Input : (32 × 32) conv1-32 (9 × 9) max-pool1 (2 × 2) conv2-32 (9 × 9) max-pool2 (4 × 4) FC-500
r_s	93.0	**93.3**	93.2	89.4
r_p	**92.4**	92.2	90.4	88.9

The kernel size also affects the performance of the network as shown in Table. 4.3. Using a greater window size than 7 × 7 leads to lower correlations, however, the network is not strongly sensitive to the kernel size when using 3 × 3, 5 × 5 and 7 × 7 especially regarding the SROCC correlation.

3.6.3 Effect of the size of input data (patches)

As mentioned earlier, the proposed CNN is fed by small patches. Since these latter are sampled in a non-overlapping way, the size affects the number of patches obtained per views (i.e smaller size leads to a larger number of patches). In this experiment, we examine how the input size affects the performance of our CNN in predicting the perceived visual quality. Table. 4.4 presents the performance of the network regarding the correlation coefficients with respect to the input patch size variation.

Tableau 3.3 – Performance of the network with respect to the input patch size variation.

Network configuration	Input : (16×16) conv1-32 (5×5) max-pool1 (2×2) conv2-32 (5×5) max-pool2 (2×2) FC-500	Input : (32×32) conv1-32 (5×5) max-pool1 (2×2) conv2-32 (5×5) max-pool2 (10×10) FC-500	Input : (64×64) conv1-32 (5×5) max-pool1 (2×2) conv2-32 (5×5) max-pool2 (26×26) FC-500	Input : (128×128) conv1-32 (5×5) max-pool1 (2×2) conv2-32 (5×5) max-pool2 (58×58) FC-500
r_s	92.7	**93.3**	92.7	89.3
r_p	90.3	**92.2**	91.8	86.8

As we can see from Table. 4.4, the performance of the network is sensitive to the size of the input patch. The best correlations are provided when using input patches with a size of 32×32. Otherwise, using patches with size 128×128 provides the lowest results since the number of patches decreases strongly and thus the learning dataset becomes smaller.

According to these experiments, we adopt the CNN configuration that leads to the best correlation scores ($r_s = 93.3\%$ and $r_p = 92.2\%$) :

- Input : (32×32)
- conv1-32 (5×5)
- max-pool1 (2×2)
- conv2-32 (5×5)
- max-pool2 (10×10)
- FC-500

3.6.4 Effect of the number of views

The 3D mesh is represented by different views obtained by fixing virtual cameras at different angles. The number of views is inversely proportional to the rotation angle of the virtual camera i.e. smaller angle provides more views. In this experiment, we test how the number of input views affects the performance of our method. Table. 4.5 presents the performance of our method on the General-purpose database using a different number of views.

Tableau 3.4 – Correlation coefficients r_s (%) and r_p (%) of SCNN-BMQA using different number of views on the General-purpose database.

Number of views (Rotation angle)	576 $(\frac{\pi}{12})$		144 $(\frac{\pi}{6})$		64 $(\frac{\pi}{4})$	
	r_s	r_p	r_s	r_p	r_s	r_p
Correlation score	90.6	90.1	93.3	92.4	88.6	87.3

It is shown in Table. 4.5 that the best performance is obtained using the angle $\frac{\pi}{6}$. Smaller angle (i.e. $\frac{\pi}{12}$) provides 576 views, although this number seems representative, many views may have the same information. Greater angle (i.e. $\frac{\pi}{4}$) provides 64 views, which is not enough to represent the 3D shape since a lot of information is missed.

3.6.5 Performances with different combination strategies

We examine in this section the performance of our method according to the different combination strategies (concatenation, element-wise multiplication, and CMP discussed in Sec. 3.5.2). We tried all the possible combinations using the three above described models that lead to four experiments by a combination type. In addition, we compute the mean correlation scores for all the databases. Table. 3.5 presents the correlation coefficients r_s (%) and r_p (%) of the fine-tuned networks on described databases with the different combinations.

TABLEAU 3.5 – Correlation coefficients r_s (%) and r_p (%) of the fine-tuned networks on LIRIS masking database, LIRIS/EPFL general-purpose database, the UWB compression database and the IEETA simplification database using different combination strategies.

Combination type	Networks	Masking		General		Compression		Simplification		Mean scores	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
No combination	VGG	92.2	91.0	91.1	88.7	85.3	84.2	84.1	83.8	88.2	86.8
	AlexNet	89.6	88.6	85.7	86.1	89.1	88.4	90.1	89.1	88.6	88.1
	ResNet	88.9	89.7	86.9	86.5	86.1	86.6	88.2	87.3	87.5	87.5
Concatenation	VGG + AlexNet	95.9	95.3	93.0	91.2	89.9	88.6	87.6	85.4	91.6	90.1
	VGG + ResNet	94.2	93.9	92.6	91.6	88.5	88.1	86.5	86.1	90.4	89.9
	AlexNet + ResNet	93.2	91.6	91.9	90.8	90.1	88.9	88.1	88.3	90.8	89.9
	All networks	**96.3**	95.1	93.6	91.9	90.5	89.3	90.0	90.2	92.6	91.6
Multiplication	VGG + AlexNet	90.1	88.9	88.1	86.9	85.2	84.2	86.3	85.6	87.4	86.4
	VGG + ResNet	91.5	90.0	87.6	88.3	84.3	84.8	89.2	88.6	88.1	87.9
	AlexNet + ResNet	89.6	90.6	88.6	87.3	86.1	84.0	84.5	83.0	87.2	86.2
	All networks	88.6	86.9	86.0	85.2	84.9	82.6	83.5	81.9	85.7	84.1
1 × 1 Convolution	All networks	90.3	90.1	91.0	89.6	85.4	86.5	86.2	85.8	88.2	88.0
	VGG + ResNet	93.7	94.3	88.9	87.8	86.7	88.8	89.7	89.0	89.8	90.0
	AlexNet + ResNet	90.6	90.4	91.0	92.6	85.8	86.3	87.5	88.9	88.7	89.5
	All networks	91.2	92.8	89.8	91.2	86.3	89.0	85.9	88.3	88.3	90.3
Compact multi-linear pooling	VGG + AlexNet	94.8	95.0	92.6	93.2	91.3	92.6	90.7	89.1	92.3	92.5
	VGG + ResNet	94.5	93.1	91.9	92.6	90.8	90.3	90.1	89.6	91.8	91.4
	AlexNet + ResNet	94.5	93.9	93.6	93.5	**93.2**	92.2	91.0	90.8	93.0	92.6
	All networks	93.2	**92.8**	**92.6**	**91.3**	90.2	**90.9**	90.6	**90.4**	**91.7**	**91.4**

In general, the three combination types provide excellent performances on all the databases as proven by the high correlation scores.

- The multiplication combination provides the lowest scores, especially when combining the three networks. Multiplying three vectors amplifies their values and leads to considerable modifications to the extracted features. Thus, the estimation is less reliable comparing to the other combinations.
- The concatenation allows obtaining an extended feature vector. This combination provides high scores and outperforms the multiplication with a considerable correlation scores improvement reaches 8.3 % (All networks r_p on the simplification database).
- 1×1 convolution performs better than the multiplication with an improvement up to 6.2 % (All networks r_p on the mean score) but not as good as the concatenation.
- The CMP combination provides the highest scores in most situations : the highest performance on the General-purpose database, the highest r_p score, and the second r_s score with a slight difference on the other databases.

From the above observations, we conclude that the best combination strategy is the CMP as proven also by the high scores when averaging over all the databases. In the following, we adopt the CMP using the three networks for the comparison with the state-of-the-art.

3.6.6 Evaluation and comparison with the state-of-the-art

In this section, we conduct a comparative study of the obtained results with the state of the art including FR (HD Aspert *et al.* (2002), RMS Cignoni *et al.* (1998) , MSDM2 Lavoué (2011), TPDM Torkhani *et al.* (2014), Yildiz et al Yildiz et Capin (2017), TPDMPW Feng *et al.* (2018), Chetouani Chetouani (2018b)), RR (3DWPM1 Corsini *et al.* (2007), 3DWPM2 Corsini *et al.* (2007), FMPD Wang *et al.* (2012), DAME Váša et Rus (2012)) and NR (NR-SVR Abouelaziz *et al.* (2016b), NR-GRNN Abouelaziz *et al.* (2016a), NR-CNN1 Abouelaziz *et al.* (2017), NR-CNN2 Abouelaziz *et al.* (2018), BMQI Nouri *et al.* (2017)). We recall that we proposed two schemes to evaluate the perceived visual quality using CNNs. In the first one we used one CNN architecture from scratch (called CNN-BMQA), and in the second scheme we combine three fine-tuned networks along with the compact bilinear pooling (called CNNs-CMP). The correlation coefficients values r_s and r_p on the LIRIS masking, LIRIS/EPFL General-purpose, UWB compression and the IEETA simplification databases are listed respectively in Tables 5.2- 5.5. The values of the state-of-the-art metrics are obtained from Lavoué (2011) for Tables 5.2- 5.4 and from Torkhani *et al.* (2012) for Table 5.5.

As shown in Tables 5.2 to 5.5, the geometric measures HD, and RMS perform the worst. One reason is that these methods do not include the main operations of the HVS and the visual quality is computed by a simple geometric distance. For the other FR measures, MSDM2 and TPDM incorporate the perceptual information, represented in the mesh curvature. As such, the perceptual information is included and better prediction is achieved compared to the geometric measures as proven by the obtained correlation coefficients. The RR method FMPD also provides good correlations compared to MSDM2 and TPDM. This method (FMPD) includes a roughness measure which is an important

Tableau 3.6 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the LIRIS/EPFL general-purpose database.

Type	Metric	Armadillo		Dyno		Venus		Rocker		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	69.5	30.2	30.9	22.6	1.6	0.8	18.1	5.5	13.8	1.3
	RMS Cignoni *et al.* (1998)	62.7	32.2	0.3	0.0	90.1	77.3	7.3	3.0	26.8	7.9
	MSDM2 Lavoué (2011)	81.6	72.8	85.9	73.5	89.3	76.5	89.6	76.1	80.4	66.2
	TPDM Torkhani *et al.* (2014)	84.5	78.8	92.2	89.0	90.6	91.0	92.2	91.4	89.6	86.2
	Yildiz et al Yildiz et Capin (2017)	-	86.0	-	79.0	-	89.0	-	88.0	-	-
	TPDMPW Feng *et al.* (2018)	-	-	-	-	-	-	-	-	87.2	87.7
	Chetouani Chetouani (2018b)	75.7	86.1	90.6	90.0	94.9	95.5	91.4	92.1	88.1	90.9
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	65.8	35.7	62.7	35.7	71.6	46.6	87.5	53.2	69.3	38.4
	3DWPM2 Corsini *et al.* (2007)	74.1	43.1	52.4	19.9	34.8	16.4	37.8	29.9	49.0	24.6
	FMPD Wang *et al.* (2012)	75.4	83.3	89.6	88.9	87.5	83.9	88.8	84.7	81.9	83.5
	DAME Váša et Rus (2012)	60.3	76.3	92.8	88.9	91.0	83.9	85.0	80.1	76.6	75.2
No-Reference	NR-SVR Abouelaziz *et al.* (2016b)	76.8	91.5	78.6	84.1	85.7	88.6	86.2	86.6	81.5	87.8
	NR-GRNN Abouelaziz *et al.* (2016a)	87.1	**97.3**	91.2	**94.1**	86.3	85.0	78.6	74.8	86.2	88.7
	NR-CNN1 Abouelaziz *et al.* (2017)	87.2	84.3	86.4	86.2	92.2	85.6	91.3	85.2	83.6	82.7
	NR-CNN2 Abouelaziz *et al.* (2018)	93.4	95.6	86.2	84.3	**94.1**	90.3	80.4	82.2	81.8	82.5
	BMQI Nouri *et al.* (2017)	20.1	-	83.5	-	88.9	-	92.7	-	78.1	-
	CNN-BMQA	90.6	92.4	88.3	86.3	93.1	92.6	88.9	89.0	**90.0**	**92.0**
	CNNs-CMP	93.4	92.9	91.6	90.9	88.9	89.4	92.6	93.9	**92.6**	**91.3**

Tableau 3.7 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the LIRIS masking database.

Type	Metric	Armadillo		Lion		Bimba		Dyno		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	48.6	37.7	71.4	25.1	25.7	7.5	48.6	31.1	26.6	4.1
	RMS Cignoni *et al.* (1998)	65.7	44.6	71.4	23.8	71.4	21.8	71.4	50.3	48.8	17.0
	MSDM2 Lavoué (2011)	88.6	65.8	94.3	87.5	**100**	93.7	**100**	91.7	89.6	76.2
	TPDM Torkhani *et al.* (2014)	88.6	91.4	82.9	88.4	**100**	97.2	**100**	**97.1**	90.0	88.6
	PDMPW Feng *et al.* (2018)	-	-	-	-	-	-	-	-	94.2	91.7
	Chetouani Chetouani (2018b)	**99.0**	**99.0**	83.0	94.0	99.0	99.0	93.0	98.0	93.9	97.8
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	58.0	41.8	20.0	9.7	20.0	8.4	66.7	45.3	29.4	10.2
	3DWPM2 Corsini *et al.* (2007)	48.6	37.9	38.3	22.0	37.1	14.4	71.4	50.1	37.4	18.2
	FMPD Wang *et al.* (2012)	94.2	88.6	93.5	94.3	98.9	**100**	96.9	94.3	80.8	80.2
	DAME Váša et Rus (2012)	94.3	96.0	**100**	**99.5**	97.7	88.0	82.9	89.4	68.1	58.6
No-Reference	NR-SVR Abouelaziz *et al.* (2016b)	89.5	84.7	**100**	96.3	94.2	93.6	94.4	89.7	90.4	91.2
	NR-GRNN Abouelaziz *et al.* (2016a)	82.3	80.5	94.1	97.0	90.2	94.3	78.2	82.3	90.2	82.4
	NR-CNN1 Abouelaziz *et al.* (2017)	95.2	97.6	89.4	91.6	93.4	98.7	96.3	89.9	88.2	85.4
	BMQI Nouri *et al.* (2017)	94.3	-	94.3	-	**100**	-	83.0	-	78.1	-
	CNN-BMQA	92.5	93.1	90.6	89.6	88.5	90.2	92.3	90.6	**91.4**	**90.8**
	CNNs-CMP	94.0	94.6	92.6	91.6	90.7	91.2	93.4	92.6	**93.2**	**92.8**

Tableau 3.8 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the UWB compression database.

Type	Metric	Bunny		James		Jessy		Nissan		Helix		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	34.1	52.2	-16.8	6.8	-23.6	12.5	14.4	23.6	45.1	46.4	10.6	28.3
	RMS Cignoni *et al.* (1998)	34.2	20.9	14.0	10.8	0.0	14.8	17.8	29.7	46.9	44.6	22.0	24.1
	MSDM2 Lavoué (2011)	**97.4**	90.1	82.6	69.2	84.3	63.1	84.4	73.1	98.1	94.7	89.3	78.0
	TPDM Torkhani *et al.* (2014)	95.1	**96.5**	90.8	73.6	85.8	75.8	82.7	73.4	**98.7**	95.0	91.5	82.9
	TPDMPW Feng *et al.* (2018)	-	-	-	-	-	-	-	-	-	-	91.3	**96.4**
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	94.7	93.4	77.3	72.3	87.2	**89.5**	63.6	59.3	98.0	**95.2**	84.1	81.9
	3DWPM2 Corsini *et al.* (2007)	96.0	91.2	76.9	65.3	86.9	85.9	56.3	67.6	95.5	94.3	82.3	80.9
	FMPD Wang *et al.* (2012)	94.2	89.6	95.3	91.2	63.3	60.0	92.4	77.5	98.4	90.8	88.8	81.8
	DAME Váša et Rus (2012)	96.8	93.4	**95.7**	**93.4**	84.4	70.5	**93.9**	75.3	96.6	**95.2**	**93.5**	85.6
No-Reference	CNN-BMQA	92.3	91.9	88.9	89.3	91.3	92.4	90.1	90.8	91.6	89.5	**90.1**	**88.3**
	CNNs-CMP	91.3	90.6	93.2	92.6	89.3	88.6	90.1	90.3	91.6	89.6	90.2	90.9

Tableau 3.9 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the IEETA simplification database.

Type	Metric	Bones		Bunny		Head		Lung		Strange		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	94.3	84.8	39.5	14.3	88.6	53.0	88.6	64.9	37.1	27.4	49.4	25.5
	RMS Cignoni *et al.* (1998)	94.3	71.1	77.1	79.2	42.9	23.1	94.3	71.3	94.3	92.4	64.3	34.4
	MSDM2 Lavoué (2011)	77.1	**96.7**	94.3	96.3	88.6	79.0	65.7	85.3	**100**	**98.1**	86.7	79.6
	TPDM Torkhani *et al.* (2014)	**99.0**	94.3	**98.0**	94.3	63.1	65.7	**98.6**	94.3	98.7	94.3	86.9	88.2
Reduced Reference	FMPD Wang *et al.* (2012)	88.6	96.0	94.3	**98.0**	65.7	70.4	88.6	**95.5**	65.7	96.0	87.2	89.3
No-Reference	CNN-BMQA	93.2	92.4	90.3	91.0	88.6	89.3	89.0	88.9	91.5	92.1	**90.4**	**90.2**
	CNNs-CMP	92.9	91.9	93.1	93.0	89.6	88.6	92.0	91.6	94.3	93.2	**90.6**	**90.4**

feature in mesh processing. The proposed method shows excellent performance on all the available subjectively-rated MVQ databases, as proven by its high scores on the individual models as well as on the whole repositories.

— The General-purpose database (see Table 5.2) is the largest MVQ database so far, it comprises the highest number of distorted meshes among all the other databases (i.e 84 distorted meshes and a variety of distortion types). On this database, the proposed method shows good performance and provide the highest correlation coefficients ($r_s = 94.4\%$ and $r_p = 94.8\%$).

— On the LIRIS masking database (see Table 5.3), our method provides the highest Spearman and Pearson correlation coefficients on the whole corpus ($r_s = 95.8\%$ and $r_p = 95.5\%$) and outperforms the NR methods (BMQI, NR-SVR, and NR-GRNN) as well as the most effective FR and RR methods.

— On the UWB compression database (see Table 5.4), the proposed method performs the best in terms of PLCC score ($r_p = 93.8\%$) outperforming the most effective methods. In addition, it provides the second higher r_s score on the whole repository

($r_s = 92.7\%$) against $r_s = 93.5\%$ for the RR method DAME.

— On the IEETA simplification database (see Table 5.5), the proposed method provides the highest correlation coefficients ($r_s = 91.0\%$ and $r_p = 91.1\%$). The perceptual methods MSDM2, TPDM, and FMPD also perform well in this database.

As we can see, our network successfully estimates the perceived visual quality using the cross dataset evaluation as proven by the high correlation coefficients obtained. These results ensure the generalization ability of SCNN-BMQA.

3.7 Conclusion

Convolutional Neural Networks differ from other forms of Artificial Neural Network in that instead of focusing on the entirety of the problem domain, knowledge about the specific type of input is exploited. This, in turn, allows for a much simpler network architecture to be set up. This chapter has outlined the basic concepts of Convolutional Neural Networks, explaining the layers required to build one and detailing how best to structure the network to be adopted in our application. We presented no-reference MVQ assessment methods to accurately estimate the perceived visual quality of distorted meshes. First, A CNN architecture from scratch is used to learn sets of 2D patches rendered from the 3D mesh. The architecture successfully predicts the visual quality of distorted meshes as proven by the high correlations with human judgment. We have tested many network configurations (the number and the size of kernels, and the size of the input data). It is demonstrated from the experiments that the CNN parameters significantly affect the performance of the network. Second, we use deeper architectures represented by fine-tuned networks (VGG, AlexNet, and Resnet). Feature vectors are first extracted using the three fine-tuned CNN models and the compact multi-linear pooling is then used to fuse the extracted feature vectors into a global feature representation. Several tests have been conducted, in particular, we test and compare different combination strategies, and we show that combining multiple DCNNs increases the performances and we can derive an effective blind MVQ method. Through comparisons with the state of the art on prominent MVQ databases, it is shown that the proposed methods provide high correlations with subjective scores and overcomes effective existing full reference and reduced reference methods. Besides, they can be useful in practical situations since they do not require any information about the reference unlike the full reference and reduced reference methods. In the next chapter, we include a very important perceptual aspect, we study the effect of visual saliency in MVQ assessment, by adopting a saliency-based patch selection before using the CNN architectures.

Chapitre

4

Mesh visual quality assessment using visual saliency

Sommaire

4.1 Introduction

We discussed previously the importance of perceptual information of 3D meshes. It is proven that perceptually-based methods estimate well the perceived visual quality compared to distance-based methods. It is thus crucial to include perceptual information to take into account the main operation of the human visual system. Visual saliency is a perceptual concept that describes the attention of our HVS to some regions due to its specificities (curvature, orientation, and so on). In this chapter, investigate the usage of CNN architectures and 3D visual saliency to estimate the perceived visual quality of distorted meshes. It is an extension of the tests performed in the last chapter, the added value is to study the effect of visual saliency in our quality estimation approach using CNNs. To do so, the CNN architecture is fed by small patches selected carefully according to their level of saliency. First, the visual saliency of the 3D mesh is computed. Afterward, we render 2D projections from the 3D mesh and its corresponding 3D saliency map. Then the obtained views are split into 2D small patches that pass through a saliency filter in order to select the most relevant patches. Finally, the CNN architectures described in the last chapter are used for feature learning and the quality score estimation. The remainder of this chapter is organized as follows : In the first section, we explain the notion of visual saliency as biological and psychological property. In the second section, we present some relevant methods of 3D visual saliency including the used method in our work. In the third section, we describe the patch selection strategy based on visual saliency and the quality score estimation. The last section is dedicated to experimental results and the study of the effect of the saliency-based patch selection.

4.2 Visual saliency as biological and psychological proprety

Visual saliency is the quality of subjective perception which makes an item stand out from its neighbors and immediately attracts our observer. Our attention is drawn to prominent visual stimuli. Therefore, the importance of the saliency of a region depends on the distinction of this region from what it surrounds. In other words, we are concerned with attentional mechanisms linked to the properties of the visual signal or stimuli rather than those associated with human observers visualizing the scene or with the task entrusted to them before visualization. Thus a visual element of the scene is said to be salient if it is easily remarkable without a priori provided to the observer or if it comes out first when viewing the scene. This element would tend to capture more attention than other elements of the scene and thus attract more attention from observers. Visual attention allows us to build a perception according to our needs and capacities. The power and the speed of this mechanism are part of the selection of the most relevant information of the scene by moving the eyes of the observer sequentially between salient regions while focusing attention. It is important for complex biological systems to quickly detect objects and regions within a given field of vision. However, simultaneously identifying all the targets of interest in a visual field causes prohibitive computational complexity, which is in fact a daunting task even for the most sophisticated biological brains, and even less for any existing computer. One solution adopted by primates and other animals is to limit the complex process of recognizing objects to a small area, or a few objects at any given time. The numerous objects or areas of the visual scene can then be processed one after the other. Visual attention can be a solution to the inability to process all places in parallel. However, this solution produces a problem, it is a question of how to choose the target of attention to treat a single object or region. Visual salience helps brains achieve effective selection in a reasonable manner. The first stages of visual processing give rise to a distinct quality of subjective perception which makes certain stimuli remarkable among other objects or locations. Our brains have evolved to quickly calculate saliency automatically and in real-time across the entire visual field. Visual attention is then drawn to the prominent visual locations.

The attentional focus towards a particular region of the visual field is influenced by two types of process : bottom-up and top-down :

— **Buttom-up** saliency is a fast, stimulus-driven process : also called ascending, is an exogenous mechanism guided only by the stimuli present in the visual field (contrast, texture, shape, ...) without any intervention on the part of the observer, The indices of the Buttom-up models are based on the characteristics of the visual scene.

— **Top-down** saliency is a slower memory-dependent process, also called descendants, is an endogenous mechanism involving the intervention of the subject, and it is determined by the (high-level) activities and visual tasks in which an organism is engaged. The indices of this model are determined by cognitive phenomena (recognition, prediction, ...). These mechanisms are linked to the task but also to the semantics of the stimuli and to the specific experience, idiosyncrasy, of each observer.

Studies of these attentional processes have shown that bottom-up mechanisms are faster and precede top-down mechanisms that are slower to implement Wolfe *et al.* (2000); Tatler *et al.* (2005); Parkhurst *et al.* (2002). The overwhelming majority of saliency models have only considered bottom-up saliency mechanisms following the fact that the bottom-up visual pathway is better understood than its top-down counterpart, in terms of both the neural circuits involved and the resulting subject behavior. We focus on our study on bottom-up processes which are only induced by the properties of the stimuli and which do not depend on the observer. These will allow us to detect perceptually protruding regions that can attract the visual attention of the observer.

Three popular components are commonly adopted. The first component, which is also the first processing stage in most saliency models, is the extraction of early visual features. These features usually include low-level simple visual attributes, such as intensity contrast, color opponency, orientation, motion, and others. The second common component of many saliency models is the adoption of a *center-surround* formulation for bottom-up saliency. It assumes that in the absence of high-level (recognition) goals, saliency is determined by how distinct the stimulus at each location of the visual field is from its surrounding. The third common practice in the design of saliency detectors is the hypothesis of the existence of a saliency map, which can be generated through either the combination of intermediate feature-specific saliency maps or the direct analysis of feature interactions.

4.3 3D mesh visual saliency : state of the art

4.3.1 State of the art

Some previous work on capturing the visual saliency of 3D meshes used saliency models defined for 2D images. These models are applied to 2D projections of 3D meshes, which are called image-based models. This approach does not take sufficient account of the relief of the 3D mesh which nevertheless constitutes a major characteristic in the perception of 3D content as mentioned in Howard et Rogers (2002).

We present in this section only model-based approaches using the geometry of the 3D mesh to predict the visual saliency. The approach of Lee *et al.* (2005) will be presented in detail afterward as it will serve as the used method in our study of the effect in visual saliency in MVQ assessment.

— In Wu *et al.* (2013), the method detects salient regions using a descriptor measuring the local height field in the neighborhood of each node ; a map of heights is generated to represent its shape. Then, the Zernike moments are extracted from these maps to obtain a rotation-invariant representation. The multi-scale aspect of the descriptor is implemented by varying the size of the height maps. Local saliency is first computed, to do so, the mesh is segmented into similar patches. The saliency of the individual nodes is obtained by interpolating the saliencies related to the neighboring patches. As for the overall saliency, a gathering of nodes into patches, which the saliency is similar, is carried out. The overall saliency for each node is then obtained by searching for the nearest patches and by interpolating their salient feature. Finally, the final visual saliency of a mode is obtained by combining and normalizing the values of global and local saliency.

— In Zhao *et al.* (2013), the authors propose a sampling-based saliency detection method for simplifying 3D meshes. The approach begins by applying a Gaussian filter on the nodes of the mesh. Then, the parameters representing the mean curvature and the directions of curvature are calculated on different scales. The different maps are filtered by a median filter before being combined to produce the final saliency map.

— In Zhao *et al.* (2012), the method detects the salient regions by diffusing the surface index using a non-local filter Buades *et al.* (2005). The patch-based approach begins by filtering the mesh to remove the high frequencies and then calculates the similarity between the nodes. Then, the mesh is transformed into volumetric data on several scales. The dissimilarity between two patches located in two subvoxels makes it possible to generate a dissimilarity map. Finally, the saliency of a patch which is proportional to its dissimilarity is defined by the average of its saliency on the different scales.

— By incorporating the CRF framework (Conditional Random Field) into an approach to compute the saliency Song *et al.* (2012), authors establish an approach generating in the first place a multi-scale representation of the mesh. This information is then combined using the CRF to label the mesh regions as salient and non-salient regions. The multi-scale representation is calculated by applying Gaussian filters on a neighborhood delimited by a geodesic radius. The Gaussian differences are calculated in each scale and represent the movement of a node after the filtering operation. These latter are then projected into the normal associated to the target node to obtain the map of a determined scale. After calculating the multi-scale representations, they are integrated into the CRF by introducing a consistency constraint between the neighboring nodes to increase the robustness of the labeling of the approach. Finally, the assignment of a label to each node of the mesh in the CRF is resolved with the Belief Propagation algorithm.

— In Zhao *et al.* (2013), the authors propose an approach to select points of interest by estimating saliency. To remove noise and make the detection of points of interest more robust, the method begins by smoothing the mesh by a bilateral filter applied to the normals of the mesh while using a relative distance instead of an absolute distance. Then the Retinex theory Elad (2005) is implemented to reinforce the local details and to estimate the invariant properties from the surface. After the segmentation of the surface, the saliency is estimated as a function of the spatial distance between the resulting segments.

- The model of Tal et al. Leifman *et al.* (2012) detects regions of interest on a 3D mesh by considering both the local and global distinction of a node as well as the edges of the shape to be analyzed. To encode the local geometry of each node, the Spin image descriptor Johnson et Hebert (1999) is used. A Spin Image is a 2D histogram reflecting the local geometry of each node by projecting its neighbors on the estimated tangent plane.
- Song et al. Song *et al.* (2014) suggest estimating the saliency in the spectral domain by analyzing the irregularities of the Log Laplacian spectrum. The authors argue that the Log-Laplacian geometry spectrum of a 3D mesh has exploitable attributes for the saliency estimation.

4.3.2 Lee method

To compute saliency for 3D meshes, Lee's method uses the center-surround operation adopted by the well known Itti et al's. method. the author considers the geometry of meshes to be the most important contributor to saliency. An overview of this method is illustrated in Fig. 4.1.

The first step consists of computing surface curvatures C and define a set of points $N(v, \sigma)$ for a vertex v within a distance σ.

$$N(v, \sigma) = \{x || x - v|| < \sigma, \text{x is a mesh point}\} \tag{4.1}$$

The Gaussian-weighted average $G(C(v), \sigma)$ of the mean curvature is computed compute as :

$$G(C(v), \sigma) = \frac{\sum_{x \in N(v, 2\sigma)} C(x) \exp[-||x - v||^2/(2\sigma^2)]}{\sum_{x \in N(v, 2\sigma)} \exp[-||x - v||^2/(2\sigma^2)]} \tag{4.2}$$

With this formulation, a cut-off for the Gaussian filter at a distance 2σ is adopted. The saliency $S(v)$ of a vertex v is computed as the absolute difference between the Gaussian-weighted averages computed at fine and coarse scales. The standard deviation for the coarse-scale as twice that of the fine-scale is used :

$$S(V) = \|G(C(v), \sigma) - G(C(v), 2\sigma)\| \tag{4.3}$$

The multi-scale saliency $S_i(V)$ of a node v on a scale i is defined by :

$$S_i(V) = \|G(C(v), \sigma_i) - G(C(v), 2\sigma_i)\| \tag{4.4}$$

where σ_i is the standard deviation of the Gaussian filter at scale i. Five scales $\sigma_i \in \{2\epsilon, 3\epsilon, 4\epsilon, 5\epsilon, 6\epsilon\}$ are used, where ϵ is 0.3% of the length of the diagonal of the bounding box of the model.

Once the saliency maps S_i associated to different scales i are obtained, a non-linear suppression operator is applied to consider only the saliency maps comprising few saliency peaks. Each saliency map S_i is normalized and the maximum saliency value M_i as well as

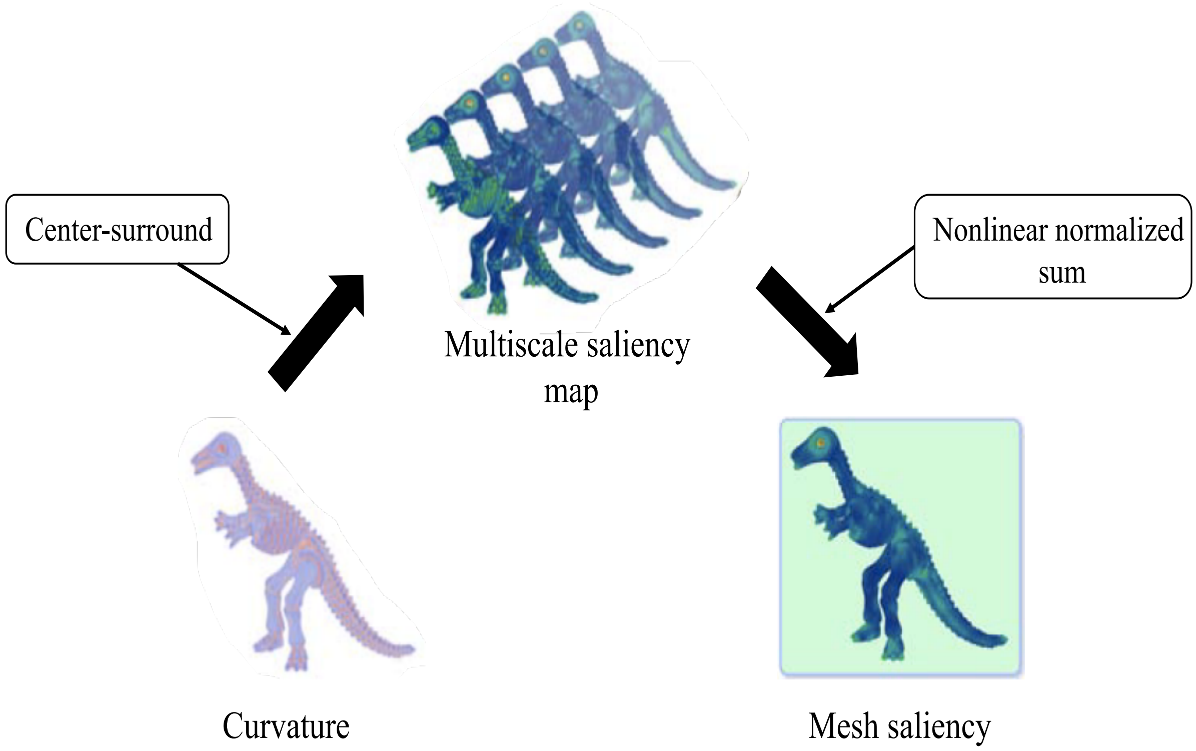


FIGURE 4.1 – Mesh Saliency Computation : The mean curvature at mesh vertices is first computed. For each vertex, saliency is computed as the difference between mean curvatures filtered with a fine and a coarse Gaussian. For each Gaussian, the Gaussian-weighted average of the curvatures is computed of vertices within a radius 2σ , where σ is Gaussian's standard deviation. The saliency at different scales is computed by varying σ. The final saliency is the aggregate of the saliency at all scales with a non-linear normalization.

the average $\tilde{m}_i$ of the local maxima are calculated. Finally, S_i is multiplied by the factor $(M_i - \tilde{m}_i)^2$ and the final saliency map S is defined by the sum of the saliency maps on all scales :

$$S = \sum_i S(f_i) \tag{4.5}$$

4.4 Introducing visual saliency in mesh quality assessment

As mentioned earlier, our attention is generally attracted by salient visual stimuli. It is important for complex biological systems to quickly detect the most relevant regions in a given field of view. Visual saliency is a subjective phenomenon that makes a region remarkable compared to others and immediately attracts our attention. The HVS has evolved to automatically detect salient regions. Visual saliency has been used previously for MVQ assessment. Anass et al. Corsini *et al.* (2007) proposed a blind mesh quality assessment index (BMQI) based on the estimation of visual saliency and roughness. In

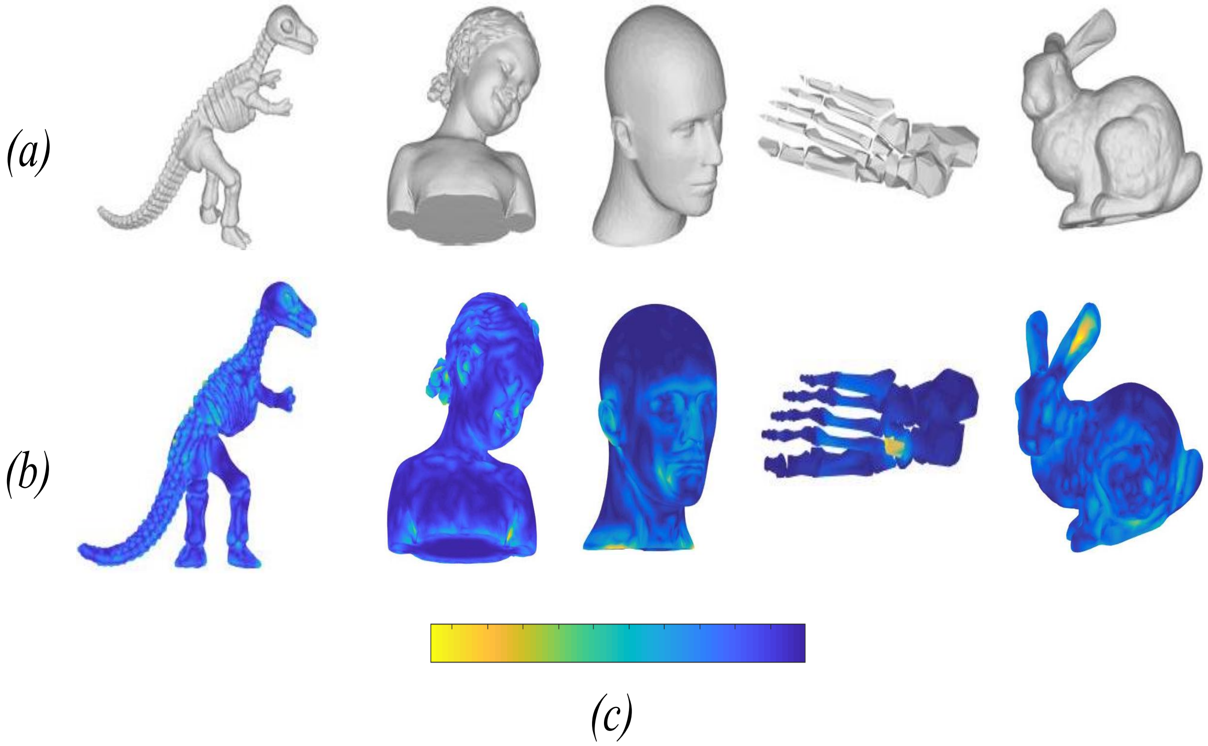


Figure 4.2 – Examples of mesh saliency : (a) 3D models, (b) their corresponding saliency maps, and (c) is the colormap. The yellow color presents the most salient regions and the blue color presents the no salient regions.

their work, they suppose that the quality of a 3D mesh is more affected in salient regions. The same authors proposed a full reference method **?** using a multiscale visual saliency map to compare the structural information between an original mesh and a distorted version. The relationship between visual saliency and distortion perception has been studied in Engelke *et al.* (2010). It is claimed that the annoyance of the distortions depends strongly on the saliency of the regions that they appear in.
Following this principle, we assume that the subjective evaluation of the visual quality of a distorted mesh is strongly related to the distortion applied to salient regions. In other words, human visual perception is impacted by the modification in salient regions since the visual attention is attracted automatically to these locations.

To study the effect of visual saliency in MVQ assessment, we include a saliency-based patch selection before using the CNN. We use mesh saliency to detect the salient patches from specific views of the 3D mesh. To compute the saliency map of 3D meshes and thus to detect perceptually important regions on mesh surfaces, we use the method proposed in Gastaldo *et al.* (2005). This method is inspired by low-level HVS operations and it is based on the center-surround mechanism adopted by the well-known Itti et al's method Babu *et al.* (2007). The process of computing mesh saliency is as follows : First, the mean curvature is computed at mesh vertices. The mean curvature is then filtered with a fine and coarse Gaussian. After that, the saliency is computed as the difference between the filtered mean curvatures within different scales by varying the Gaussian's standard deviation. The final saliency map is obtained by a non-linear normalization sum of all the multiscale saliency maps. Fig. 4.2. shows some examples of 3D meshes and their saliency maps. We can notice that some regions in the 3D shape are considerably more distinct and hence judged as salient. Remarkably, regions with high curvature levels such as ears, nose, eyes, and paws attract more attention compared to smooth regions where the level of curvature is low.

It is noteworthy that in our work we do not propose a saliency method, we are decided to use an existing one. This method was evaluated by the authors and the obtained results validate its good performance in capturing the salient regions. Our purpose in using visual saliency is to demonstrate its importance and usefulness in quality assessment, and show how the salient regions are more susceptible to degradations that are easily detected by the human eye compared to the non-salient regions. Once the saliency map is obtained, we render 2D views following the same procedure described in the last section. The views obtained from the saliency map are used to select the salient locations in the corresponding 3D mesh as follows :

— First, we sample non-overlapping patches of size 32 × 32 from the 2D projections of the 3D mesh and its corresponding saliency map.

— For each patch of the saliency map, we compute the local level of saliency (LoS) which corresponds here to its average saliency value. The level of saliency LoS is used afterward to select the most relevant patches with a saliency threshold S_t set experimentally to 0.4. All the patches with $LoS \geq S_t$ are considered as relevant regions, whereas patches with $LoS < S_t$ are neglected. We note that the LoS is computed using only the pixels that contain the saliency information, the background pixels at object boundary are not considered. Thus, informative patches (with high saliency) at object boundaries are not ignored.

— After that, we perform a local normalization on the retrieved patches which correspond to the salient regions in the 3D mesh.

— Finally, the selected patches are then used as input to our CNN models. We use in this chapter the results obtained by the CNN from scratch as well as the combination of the fine-tuned networks presented in the last chapter.

4.5 Expiremtal results

We tested in the last chapter several parameters in order to investigate how the performance is affected and choose the best configuration, we arrived that the best configuration is : Input (32 × 32), conv1-32 (5 × 5), max-pool1 (2 × 2), conv2-32 (5 × 5), max-pool2 (10 × 10), FC-500. We also tested fine-tuned networks with different combination strategies, we arrived that the best results are obtained by using Alexnet, Resnet, and VGG with the compact multi-linear pooling.

In this section, we use the same configurations, however, this time we include the saliency-based patch selection to investigate the importance of visual saliency in predicting the perceived visual quality. Besides, we conduct the same experiments presented in the last chapter to test the effect of the network parameters using the saliency data. Afterward, we compare the results with the state of the art methods.

4.5.1 Effect of the saliency-based patch selection

As mentioned earlier, the patch selection strategy is based on the mesh saliency obtained from the distorted meshes. The patches are selected by fixing a saliency threshold. To demonstrate the importance of the patch selection strategy used in our method, we conduct an experiment by testing the ability of the proposed CNN architecture and the fine-tuned networks to predict the perceived visual quality with different saliency thresholds on the General-purpose database. In addition, we compare the performance of the networks with and without using the selection strategy ($S_t = 0$) on the four databases. Fig. 4.3 presents the correlation coefficients obtained by the trained networks with different saliency thresholds on the General-purpose database. We note that the tests are conducted only on this database since it contains a greater number of meshes. After that, the saliency threshold is generalized for the other databases.

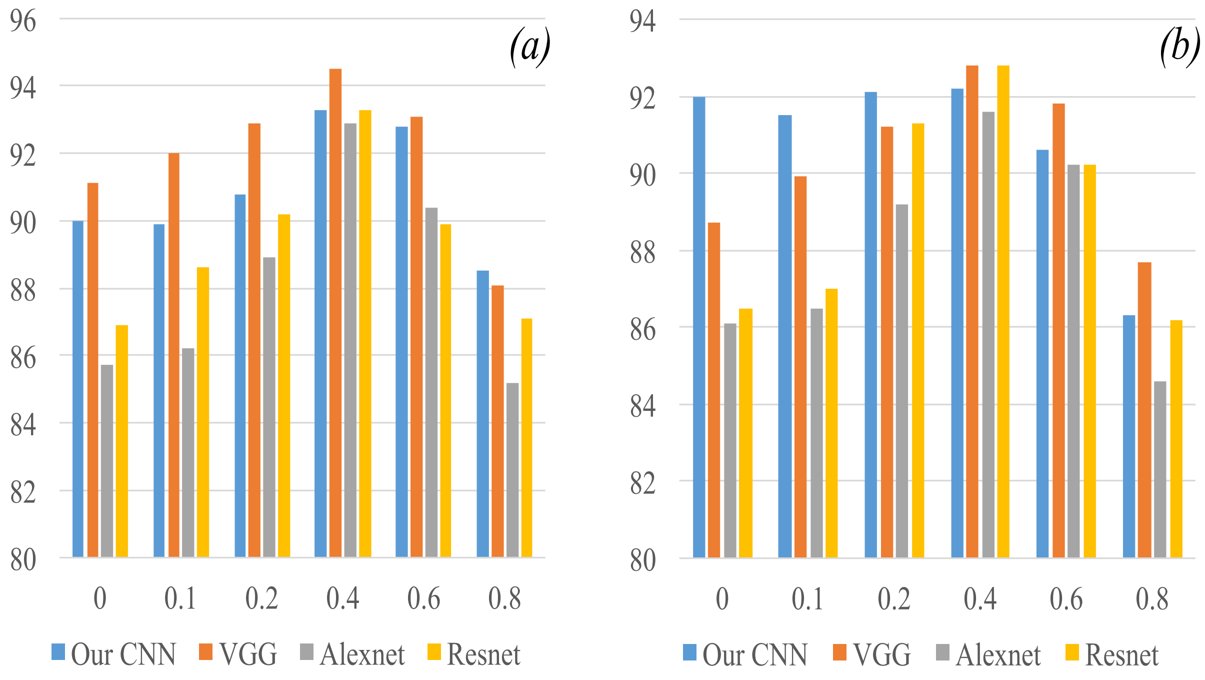


FIGURE 4.3 – Correlation coefficients r_s (%) (a) and r_p (%) (b) of the proposed CNN and the fine-tuned networks using different saliency threshold with patch size 32 × 32 on the LIRIS/EPFL general-purpose database..

When $S_t = 0.4$, the correlation scores obtain their best value for all the networks. However, when $S_t > 0.4$, the number of selected patches decreases clearly, especially when $S_t = 0.8$ (only patches with a saliency superior to 0.8 are selected). Thus, the performance, in this case, is the worst. When $S_t < 0.4$, the correlation scores decrease since the selection is less important in this case and this refers to the influence of the patch selection strategy in our method. Comparing to $S_t = 0$ (without selection), it is remarkable that the patch selection strategy significantly improves the performance as proven by the difference between the score regarding $S_t = 0$ and $S_t = 0.4$ (best saliency threshold). The saliency threshold is thus fixed to 0.4

The importance of the saliency selection strategy is confirmed in Table 4.1 that presents the averaged scores obtained by the networks with and without the selection over all databases.

Tableau 4.1 – Correlation coefficients r_s (%) and r_p (%) of the networks with and without the patch selection strategy on the four tested databases.

		Masking		General-Purpose		Compression		Simplification	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Without patch selection	Our CNN	91.4	90.8	90.0	92.0	90.1	88.3	90.4	90.2
	VGG	92.2	91.0	91.1	88.7	85.3	84.2	84.1	83.8
	AlexNet	89.6	88.6	85.7	86.1	89.1	88.4	90.1	89.1
	ResNet	88.9	89.7	86.9	86.5	86.1	86.6	88.2	87.3
With patch selection	Our CNN	95.5	94.3	93.3	92.4	90.4	88.2	90.4	90.5
	VGG	96.2	94.2	94.5	92.8	89.5	86.7	89.8	88.3
	AlexNet	95.4	93.3	92.9	91.6	92.2	91.4	92.5	91.4
	ResNet	93.4	92.2	93.3	92.8	90.4	89.9	91.6	90.4

The correlation scores increase remarkably when using saliency selection. For all networks in all databases, the improvement is between 2.3 % (AlexNet r_p on the simplification database) and 7.2 % (AlexNet r_s on the general database).
Concerning the proposed CNN the patch selection process improves significantly the correlations scores. Especially on the masking database where the Spearman coefficient increases by 4.1% and the Pearson coefficients increases by 3.5%, and on the General-purpose database where r_s and r_p coefficients increase by 3.3% and 0.9% respectively. On the UWB compression and simplification databases, the performance is slightly improved except for the Pearson correlation on the compression database which is decreased by 0.1%.
From these results, we conclude that the used patch selection strategy based on visual saliency is very effective, especially on the LIRIS masking and the General-purpose databases.

4.5.2 Effect of the number of filters in the convolutional layers

In this section, we test the ability of our network in predicting the visual quality by using a variety of convolution kernels while fixing the other parameters. Table. 4.2 presents the performance of the network regarding the correlation coefficients with respect to the size of convolution kernels.

It is shown from Table. 4.2 that the number of kernels significantly affects the performance of the network. Using 32 kernels instead of 10 leads to an important improvement, however, using more kernels than 32 decreases the performance of the network in predicting the visual quality.

Tableau 4.2 – Performance of the network with respect to the number of kernels.

Network configuration	Input : (32 × 32) conv1-10 (5 × 5) max-pool1 (2 × 2) conv2-10 (5 × 5) max-pool2 (10 × 10) FC-500	Input : (32 × 32) conv1-32 (5 × 5) max-pool1 (2 × 2) conv2-32 (5 × 5) max-pool2 (10 × 10) FC-500	Input : (32 × 32) conv1-50 (5 × 5) max-pool1 (2 × 2) conv2-50 (5 × 5) max-pool2 (10 × 10) FC-500
r_s	88.6	**93.3**	92.4
r_p	89.7	**92.2**	91.2

4.5.3 Effect of the size of filters

Now, the input patch size and the number of convolution kernels are fixed, and we test different sizes of the kernels. Table. 4.3 presents the performance of the network regarding the correlation coefficients with respect to the number of convolution kernels.

Tableau 4.3 – Performance of the network with respect to the size of convolutional kernels.

Network configuration	Input : (32 × 32) conv1-32 (3 × 3) max-pool1 (2 × 2) conv2-32 (3 × 3) max-pool2 (13 × 13) FC-500	Input : (32 × 32) conv1-32 (5 × 5) max-pool1 (2 × 2) conv2-32 (5 × 5) max-pool2 (10 × 10) FC-500	Input : (32 × 32) conv1-32 (7 × 7) max-pool1 (2 × 2) conv2-32 (7 × 7) max-pool2 (7 × 7) FC-500	Input : (32 × 32) conv1-32 (9 × 9) max-pool1 (2 × 2) conv2-32 (9 × 9) max-pool2 (4 × 4) FC-500
r_s	93.0	**93.3**	93.2	89.4
r_p	**92.4**	92.2	90.4	88.9

Using a greater window size than 7 × 7 leads to lower correlations, however, the network is not strongly sensitive to the kernel size when using 3 × 3, 5 × 5 and 7 × 7 especially regarding the SROCC correlation.

4.5.4 Effect of the size of input data (patches)

In this experiment, we examine how the input size affects the performance of our CNN in predicting the perceived visual quality. Table. 4.4 presents the performance of the network regarding the correlation coefficients with respect to the input patch size variation.

The performance of the network is sensitive to the size of the input patch. The best correlations are provided when using input patches with a size of 32 × 32. Otherwise, using patches with size 128 × 128 provides the lowest results since the number of patches

Tableau 4.4 – Performance of the network with respect to the input patch size variation.

Network configuration	Input : (16 × 16) conv1-32 (5 × 5) max-pool1 (2 × 2) conv2-32 (5 × 5) max-pool2 (2 × 2) FC-500	Input : (32 × 32) conv1-32 (5 × 5) max-pool1 (2 × 2) conv2-32 (5 × 5) max-pool2 (10 × 10) FC-500	Input : (64 × 64) conv1-32 (5 × 5) max-pool1 (2 × 2) conv2-32 (5 × 5) max-pool2 (26 × 26) FC-500	Input : (128 × 128) conv1-32 (5 × 5) max-pool1 (2 × 2) conv2-32 (5 × 5) max-pool2 (58 × 58) FC-500
r_s	92.7	**93.3**	92.7	89.3
r_p	90.3	**92.2**	91.8	86.8

decreases strongly and thus the learning dataset becomes smaller.

The results using saliency data with different network configuration confirms that the best CNN configuration that leads to the best correlation scores ($r_s = 93.3\%$ and $r_p = 92.2\%$) :

- Input : (32 × 32)
- conv1-32 (5 × 5)
- max-pool1 (2 × 2)
- conv2-32 (5 × 5)
- max-pool2 (10 × 10)
- FC-500

4.5.5 Effect of the number of views

In this experiment, we test how the number of input views affects the performance of the network. Table. 4.5 presents the performance of our method on the General-purpose database using a different number of views.

Tableau 4.5 – Correlation coefficients r_s (%) and r_p (%) of SCNN-BMQA using different number of views on the General-purpose database.

Number of views (Rotation angle)	576 ($\frac{\pi}{12}$)		144 ($\frac{\pi}{6}$)		64 ($\frac{\pi}{4}$)	
	r_s	r_p	r_s	r_p	r_s	r_p
Correlation score	90.6	90.1	93.3	92.4	88.6	87.3

It is shown in Table. 4.5 that the best performance is obtained using the angle $\frac{\pi}{6}$. Smaller angle (i.e. $\frac{\pi}{12}$) provides 576 views, although this number seems representative, many views may have the same information. Greater angle (i.e. $\frac{\pi}{4}$) provides 64 views, which is not enough to represent the 3D shape since a lot of information is missed.

4.6 Comparison with state of the art

In this section, we compare the proposed MVQ methods including visual saliency with existing methods. In the first one we used one CNN architecture from scratch (called SCNN-BMQA), and in the second scheme, we combine three fine-tuned networks along with the compact bilinear pooling (called SCNNs-CMP). The correlation coefficients values r_s and r_p on the LIRIS masking, LIRIS/EPFL General-purpose, UWB compression and the IEETA simplification databases are listed in Tables 5.2- 5.3- 5.4- 5.5

Tableau 4.6 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the LIRIS/EPFL general-purpose database.

Type	Metric	Armadillo		Dyno		Venus		Rocker		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	69.5	30.2	30.9	22.6	1.6	0.8	18.1	5.5	13.8	1.3
	RMS Cignoni *et al.* (1998)	62.7	32.2	0.3	0.0	90.1	77.3	7.3	3.0	26.8	7.9
	MSDM2 Lavoué (2011)	81.6	72.8	85.9	73.5	89.3	76.5	89.6	76.1	80.4	66.2
	TPDM Torkhani *et al.* (2014)	84.5	78.8	92.2	89.0	90.6	91.0	92.2	91.4	89.6	86.2
	Yildiz et al Yildiz et Capin (2017)	-	86.0	-	79.0	-	89.0	-	88.0	-	-
	TPDMPW Feng *et al.* (2018)	-	-	-	-	-	-	-	-	87.2	87.7
	Chetouani Chetouani (2018b)	75.7	86.1	90.6	90.0	94.9	95.5	91.4	92.1	88.1	90.9
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	65.8	35.7	62.7	35.7	71.6	46.6	87.5	53.2	69.3	38.4
	3DWPM2 Corsini *et al.* (2007)	74.1	43.1	52.4	19.9	34.8	16.4	37.8	29.9	49.0	24.6
	FMPD Wang *et al.* (2012)	75.4	83.3	89.6	88.9	87.5	83.9	88.8	84.7	81.9	83.5
	DAME Váša et Rus (2012)	60.3	76.3	92.8	88.9	91.0	83.9	85.0	80.1	76.6	75.2
No-Reference	NR-SVR Abouelaziz *et al.* (2016b)	76.8	91.5	78.6	84.1	85.7	88.6	86.2	86.6	81.5	87.8
	NR-GRNN Abouelaziz *et al.* (2016a)	87.1	**97.3**	91.2	**94.1**	86.3	85.0	78.6	74.8	86.2	88.7
	NR-CNN1 Abouelaziz *et al.* (2017)	87.2	84.3	86.4	86.2	92.2	85.6	91.3	85.2	83.6	82.7
	NR-CNN2 Abouelaziz *et al.* (2018)	93.4	95.6	86.2	84.3	**94.1**	90.3	80.4	82.2	81.8	82.5
	BMQI Nouri *et al.* (2017)	20.1	-	83.5	-	88.9	-	92.7	-	78.1	-
	CNN-BMQA	90.6	92.4	88.3	86.3	93.1	92.6	88.9	89.0	**90.0**	**92.0**
	CNNs-CMP	93.4	92.9	91.6	90.9	88.9	89.4	92.6	93.9	**92.6**	**91.3**
	SCNN-BMQA	89.8	91.4	91.6	92.2	94.6	93.8	91.9	93.9	93.3	92.4
	SCNNs-CMP	**95.8**	95.6	**93.6**	92.9	93.4	**91.3**	**94.5**	**95.2**	94.4	94.8

From these tables, the proposed methods CNN-BMQA, CNNs-CMP (without saliency) and SCNN-BMQA, SCNNs-CMP (with saliency) provides good results in general. We can easily conclude that the methods that include visual saliency perform better as proven by the remarkable differences concerning the correlation scores.

4.6.1 Psychometric curve fitting

Since the objective scores obtained by an MVQ method and the corresponding subjective scores are non-linear, it is important to introduce a psychometric fitting function to partially remove the non-linearity and make the correlation values interpretable by users. In our work, we use the cumulative Gaussian psychometric function Lavoué (2009)

Tableau 4.7 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the LIRIS masking database.

Type	Metric	Armadillo		Lion		Bimba		Dyno		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	48.6	37.7	71.4	25.1	25.7	7.5	48.6	31.1	26.6	4.1
	RMS Cignoni *et al.* (1998)	65.7	44.6	71.4	23.8	71.4	21.8	71.4	50.3	48.8	17.0
	MSDM2 Lavoué (2011)	88.6	65.8	94.3	87.5	**100**	93.7	**100**	91.7	89.6	76.2
	TPDM Torkhani *et al.* (2014)	88.6	91.4	82.9	88.4	**100**	97.2	**100**	**97.1**	90.0	88.6
	PDMPW Feng *et al.* (2018)	-	-	-	-	-	-	-	-	94.2	91.7
	Chetouani Chetouani (2018b)	**99.0**	**99.0**	83.0	94.0	99.0	99.0	93.0	98.0	93.9	97.8
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	58.0	41.8	20.0	9.7	20.0	8.4	66.7	45.3	29.4	10.2
	3DWPM2 Corsini *et al.* (2007)	48.6	37.9	38.3	22.0	37.1	14.4	71.4	50.1	37.4	18.2
	FMPD Wang *et al.* (2012)	94.2	88.6	93.5	94.3	98.9	**100**	96.9	94.3	80.8	80.2
	DAME Váša et Rus (2012)	94.3	96.0	**100**	**99.5**	97.7	88.0	82.9	89.4	68.1	58.6
No-Reference	NR-SVR Abouelaziz *et al.* (2016b)	89.5	84.7	**100**	96.3	94.2	93.6	94.4	89.7	90.4	91.2
	NR-GRNN Abouelaziz *et al.* (2016a)	82.3	80.5	94.1	97.0	90.2	94.3	78.2	82.3	90.2	82.4
	NR-CNN1 Abouelaziz *et al.* (2017)	95.2	97.6	89.4	91.6	93.4	98.7	96.3	89.9	88.2	85.4
	BMQI Nouri *et al.* (2017)	94.3	-	94.3	-	**100**	-	83.0	-	78.1	-
	CNN-BMQA	92.5	93.1	90.6	89.6	88.5	90.2	92.3	90.6	**91.4**	**90.8**
	CNNs-CMP	94.0	94.6	92.6	91.6	90.7	91.2	93.4	92.6	**93.2**	**92.8**
	SCNN-BMQA	92.4	91.7	92.2	93.1	97.9	97.3	93.4	92.6	95.5	94.3
	SCNNs-CMP	96.2	95.5	93.1	92.4	92.5	92.8	94.2	94.0	95.8	95.9

Tableau 4.8 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the UWB compression database.

Type	Metric	Bunny		James		Jessy		Nissan		Helix		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	34.1	52.2	-16.8	6.8	-23.6	12.5	14.4	23.6	45.1	46.4	10.6	28.3
	RMS Cignoni *et al.* (1998)	34.2	20.9	14.0	10.8	0.0	14.8	17.8	29.7	46.9	44.6	22.0	24.1
	MSDM2 Lavoué (2011)	**97.4**	90.1	82.6	69.2	84.3	63.1	84.4	73.1	98.1	94.7	89.3	78.0
	TPDM Torkhani *et al.* (2014)	95.1	**96.5**	90.8	73.6	85.8	75.8	82.7	73.4	**98.7**	95.0	91.5	82.9
	TPDMPW Feng *et al.* (2018)	-	-	-	-	-	-	-	-	-	-	91.3	**96.4**
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	94.7	93.4	77.3	72.3	87.2	**89.5**	63.6	59.3	98.0	**95.2**	84.1	81.9
	3DWPM2 Corsini *et al.* (2007)	96.0	91.2	76.9	65.3	86.9	85.9	56.3	67.6	95.5	94.3	82.3	80.9
	FMPD Wang *et al.* (2012)	94.2	89.6	95.3	91.2	63.3	60.0	92.4	77.5	98.4	90.8	88.8	81.8
	DAME Váša et Rus (2012)	96.8	93.4	**95.7**	**93.4**	84.4	70.5	**93.9**	75.3	96.6	**95.2**	**93.5**	85.6
No-Reference	CNN-BMQA	92.3	91.9	88.9	89.3	91.3	92.4	90.1	90.8	91.6	89.5	**90.1**	**88.3**
	CNNs-CMP	91.3	90.6	93.2	92.6	89.3	88.6	90.1	90.3	91.6	89.6	90.2	90.9
	SCNN-BMQA	95.8	91.7	96.2	95.6	92.3	90.5	88.7	84.7	96.7	94.6	90.4	88.2
	SCNNs-CMP	95.6	94.8	**92.5**	90.6	92.5	87.1	88.7	**89.0**	90.7	90.4	92.7	93.8

adopted by Bian *et al.* (2009); Vapnik (2013) defined as follows :

$$p(a, b, X) = \frac{1}{\sqrt{2\pi}} \int_{a+bX}^{\propto} \exp - \left(\frac{t^2}{2}\right) dt \tag{4.6}$$

Tableau 4.9 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the IEETA simplification database.

Type	Metric	Bones		Bunny		Head		Lung		Strange		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	94.3	84.8	39.5	14.3	88.6	53.0	88.6	64.9	37.1	27.4	49.4	25.5
	RMS Cignoni *et al.* (1998)	94.3	71.1	77.1	79.2	42.9	23.1	94.3	71.3	94.3	92.4	64.3	34.4
	MSDM2 Lavoué (2011)	77.1	**96.7**	94.3	96.3	88.6	79.0	65.7	85.3	**100**	**98.1**	86.7	79.6
	TPDM Torkhani *et al.* (2014)	**99.0**	94.3	**98.0**	94.3	63.1	65.7	**98.6**	94.3	98.7	94.3	86.9	88.2
Reduced Reference	FMPD Wang *et al.* (2012)	88.6	96.0	94.3	**98.0**	65.7	70.4	88.6	**95.5**	65.7	96.0	87.2	89.3
No-Reference	CNN-BMQA	93.2	92.4	90.3	91.0	88.6	89.3	89.0	88.9	91.5	92.1	**90.4**	**90.2**
	CNNs-CMP	92.9	91.9	93.1	93.0	89.6	88.6	92.0	91.6	94.3	93.2	**90.6**	**90.4**
	SCNN-BMQA	96.8	93.7	96.7	90.9	96.4	93.6	94.3	89.6	95.4	93.5	90.4	90.5
	SCNNs-CMP	91.3	88.9	91.1	92.4	**91.8**	**91.5**	95.3	89.4	91.1	88.9	91.0	91.1

where X is the quality score obtained by the objective method, a and b are two parameters to be determined. The values of a and b are retrieved using the MOS values and the objective values obtained by SCNN-BMQA for each database. Fig. 4.4 shows the scatter plots of the predicted scores obtained by SCNN-BMQA and the subjective MOSs. As illustrated by this figure, the subjective vs objective scores point cloud is close enough to the psychometric curve concerning the four tested databases. The good fitting of these plots is another indicator of the good performance of SCNN-BMQA.

4.6.2 Cross dataset evaluation

In this section, we investigate the generalization ability of SCNN-BMQA and SCNNs-CMP . To do so, we perform a cross-database evaluation by training the networks on the General-purpose database and using the other databases for the test. We choose this database for the training process because it contains the highest number of distorted models and rich variety of distortion types. Tables. 5.6 and 4.11 show the results of the cross dataset evaluation with SCNN-BMQA and SCNNs-CMP respectively. These tables present the correlation coefficients of each 3D object in the three tested databases (i.e LIRIS masking, UWB compression, and IEETA simplification) as well as the scores for the whole repositories.

4.7 Conclusion

Visual saliency is a perceptual feature representing an important characteristic of the human visual system. this property is crucial to explore the surrounding visual world and it is considered one of the most important elements in human perception. MVQ methods require perceptual information to include the properties of the HVS. We presented in this chapter an extension of the methods based on CNN architectures presented in the last chapter. We relied on the assumption that the subjective evaluation of the visual quality of a distorted mesh is strongly related to the distortion applied to salient regions. To demonstrate this, we used a patch selection approach based on mesh visual saliency and

Tableau 4.10 – Correlation coefficients r_s (%) and r_p (%) of SCNN-BMQA obtained by training on the General-purpose and testing on LIRIS masking, UWB compression and IEETA simplification.

Database	Object	Scores		Database	Object	Scores		Database	Object	Scores	
		r_s	r_p			r_s	r_p			r_s	r_p
LIRIS Maskig	Armadillo	88.6	86.5	UWB Compression	Bunny	89.6	88.8	IEETA Simplification	Bones	82.6	82.9
	Lion	89.5	88.8		James	86.4	84.7		Bunny	91.1	90.8
	Bimba	94.8	94.3		Jessy	82.9	82.5		Head	80.6	79.6
	Dyno	92.9	91.2		Nissan	90.4	88.7		Lung	82.4	83.8
					Helix	84.1	83.0		Strange	84.2	83.1
	All models	91.3	90.5		All models	83.7	82.8		All models	81.8	81.3

Tableau 4.11 – Correlation coefficients r_s (%) and r_p (%) of SCNNs-CMP obtained by training on the General-purpose and testing on LIRIS masking, UWB compression and IEETA simplification.

Database	Object	Scores		Database	Object	Scores		Database	Object	Scores	
		r_s	r_p			r_s	r_p			r_s	r_p
LIRIS Maskig	Armadillo	90.1	89.1	UWB Compression	Bunny	91.7	90.4	IEETA Simplification	Bones	86.3	85.8
	Lion	92.9	93.1		James	89.5	89.5		Bunny	93.4	91.2
	Bimba	96.3	95.9		Jessy	86.8	85.7		Head	84.6	82.3
	Dyno	95.3	93.6		Nissan	92.4	91.2		Lung	86.4	85.9
					Helix	88.5	86.4		Strange	85.4	84.8
	All models	94.8	93.4		All models	87.6	85.9		All models	86.4	83.2

only the salient patches are used to feed the CNN architectures. Several thresholds are tested before selecting the best one. It is proven that the used patch selection strategy based on visual saliency is very effective according to the remarkable improvement of the correlation scores. We can conclude that the distortions in salient regions are more important than in the normal regions, and thus, the saliency information impacts more the overall subjective score.

The current stage of development so far is to use CNN fed by 2D patches so-called view-depending methods. To process directly the 3D mesh, we introduce in the next chapter a model-based method using graph convolutional networks (GCN) to work directly on the 3D model itself.

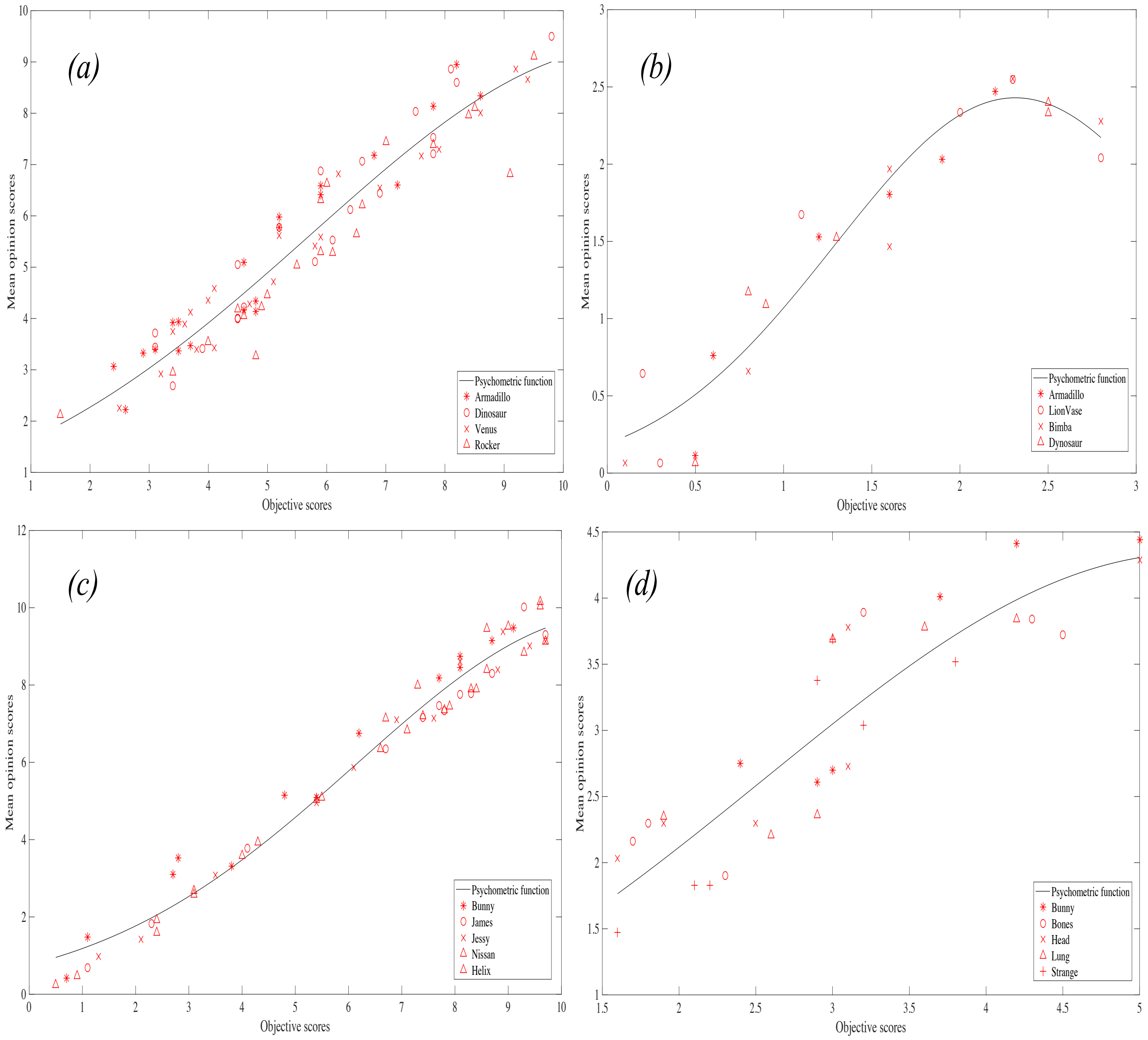


FIGURE 4.4 – Scatter plots of the mean opinion scores (MOS) versus the objective scores obtained from SCNN-BMQA. (a) LIRIS/EPFL general-purpose database. (b) LIRIS masking database. (c) the UWB compression database. (d) IEETA simplification database.

Chapitre

5

Graph convolutional neural network for mesh quality assessment

Sommaire

5.1 Introduction

In the previous chapter, we investigated the use of deep convolutional neural networks to estimate the perceived visual quality of distorted 3D meshes. We used a CNN architecture from scratch and fine-tuned networks with a specific combination for the final score. We render 2D images from multiple views of the 3D mesh. Then, each image is split into small patches which are learned to convolutional neural networks. The 3D visual saliency is adopted to select the most relevant regions with high saliency level before using the selected patches as input for the convolutional neural network. These methods show good performances and successfully predict the perceived visual quality. However, the perceptual mechanisms are applied to the images generated from the 3D data making the evaluation view-dependant also called an image-based approach. As demonstrated by Rogowitz and Rushmeier Rogowitz et Rushmeier (2001), the main problem of the image-based metrics is that, in general, the perceived degradation of still images may not be adequate to evaluate the perceived degradation of the equivalent 3D model. To overcome this limitation, we introduce a model-based method using graph convolutional networks (GCN) to work directly on the 3D model itself. GCNs has been mainly applied for the node classification tasks in which the convolution representation vector for a node functions as the only features to classify that node. For the mesh quality assessment task, the 3D model is represented by a graph and the graph-based convolution vector of nodes can be used to predict the perceived visual quality.

In this chapter, we present a new approach to estimate the perceived quality of 3D meshes. The method is no-reference and it is based on a Graph convolutional network with three modules (convolution, pooling, and classification) to rely on the problem of node classification and estimate the perceived quality score.

In the first section of this chapter, we describe the different components of our mesh visual quality assessment method including the different modules of the used GCN. In the second section, we present the training and classification protocol used for node classification. In the fourth section, we evaluate the performance of our perceptual objective measures on four subjective databases publicly available.

5.2 Background

5.2.1 From convolutional neural networks to graph convolutional networks

Convolutional neural networks offer an efficient architecture to extract highly meaningful statistical patterns in large-scale and high-dimensional datasets. The ability of CNNs to learn local stationary structures and compose them to form multi-scale hierarchical patterns has led to breakthroughs in image, video, and sound recognition tasks LeCun *et al.* (2015). Precisely, CNNs extract the local stationarity property of the input data or signals by revealing local features that are shared across the data domain. These similar features are identified with localized convolutional filters or kernels, which are learned from the data. Convolutional filters are shift- or translation-invariant filters, meaning they can recognize identical features independently of their spatial locations. Localized kernels or compactly supported filters refer to filters that extract local features independently of the input data size, with a support size that can be much smaller than the input size.

CNNs have recently attracted the attention of many researchers. They have been successfully employed in various computer vision applications allowing them to reach high performances. One of their main advantage over classical neural networks is that they adequately consider the spatial structure of the input data. Moreover, CNN allows the important property of weights sharing between the convolutional layers which restrict the number of parameters to learn. In quality assessment, the use of CNNs has shown notable

improvement in terms of the correlation with human judgment.
Some data sets can naturally be modeled as scalar functions defined on the vertices of graphs. For example, computer networks, transportation (road, rail, airplane) networks or social networks can all be described by graphs, with the vertices corresponding to individual computers, cities, or people respectively. These data can be structured with graphs, which are universal representations of heterogeneous pairwise relationships. Graphs can encode complex geometric structures and can be studied with strong mathematical tools such as spectral graph theory Chung et Graham (1997).
A generalization of CNNs to graphs is not straightforward as the convolution and pooling operators are only defined for regular grids. This makes this extension challenging, both theoretically and implementation-wise. The major bottleneck of generalizing CNNs to graphs is the definition of localized graph filters which are efficient to evaluate and learn. In the next, We provide a brief introduction to the required background in graph theory.

5.2.2 Notions for graph theory

Graphs : A graph G is a pair $(V;E)$ with $V = \{v_1, ..., v_n\}$ the set of vertices and $E \subseteq V \times V$ the set of edges. Let n be the number of vertices and m the number of edges. Each graph can be represented by an adjacency matrix A of size $n \times n$, where $A_{i,j} = 1$ if there is an edge from vertex v_i to vertex v_j , and $A_{i,j} = 0$ otherwise. In this case, we say that vertex v_i has position i in A. Moreover, if $A_{i,j} = 1$ we say v_i and v_j are adjacent. Node and edge attributes are features that attain one value for each node and edge of a graph. We use the term attribute value instead of label to avoid confusion with the graph-theoretical concept of a labeling. A walk is a sequence of nodes in a graph, in which consecutive nodes are connected by an edge. A path is a walk with distinct nodes. We write $d(u;v)$ to denote the distance between $u$and v, that is, the length of the shortest path between u and v. $N_1(v)$ is the 1-neighborhood of a node, that is, all nodes that are adjacent to v.

weighted graphs : A weighted graph $G = \{E, V, \omega\}$ consists of a set of vertices V , a set of edges E, and a weight function $\omega : E \rightarrow \mathbb{R}^+$ which assigns a positive weight to each edge. We consider here only finite graphs where $|V| = N < \infty$. The adjacency matrix A for a weighted graph G is the $N \times N$ matrix with entries $a_{m,n}$ where

$$a_{m,n} = \begin{cases} \omega(e) \text{ if } e \in E \text{ connects vertices m and n} \\ 0 \text{ otherwise} \end{cases}$$

In the present work, we consider only undirected graphs, which correspond to symmetric adjacency matrices. We do not consider the possibility of negative weights. A graph is said to have loops if it contains edges that connect a single vertex to itself. Loops imply the presence of nonzero diagonal entries in the adjacency matrix. For a weighted graph, the degree of each vertex m, written as $d(m)$, is defined as the sum of the weights of all the edges incident to it. This implies $d(m) = \sum_n a_{m,n}$ define the matrix D to have diagonal elements equal to the degrees, and zeros elsewhere.

Every real-valued function $f : V \to \mathbb{R}$ on the vertices of the graph G can be viewed as a vector in $\mathbb{R}^N$, where the value of f on each vertex defines each coordinate. This implies an implicit numbering of the vertices. We adopt this identification and will write $f \in \mathbb{R}^N$ for functions on the vertices of the graph, and $f(m)$ for the value on the m^{th} vertex.
Weighted graphs provide a flexible generalization of regular grid domains. By identifying the grid points with vertices and connecting adjacent grid points with edges with weights inversely proportional to the square of the distance between neighbors, a regular lattice can be represented with a weighted graph. A general weighted graph, however, has no restriction on the regularity of vertices. For example points on the original lattice may be removed, yielding a ”damaged grid”, or placed at arbitrary locations corresponding to irregular sampling. In both of these cases, a weighted graph can still be constructed that represents the local connectivity of the underlying data points.
Similarly, weighted graphs can be inferred from mesh descriptions for geometrical domains. An enormous literature exists on techniques for generating and manipulating meshes, such structures are widely used in applications for computer graphics and numerical solution of partial differential equations.
Weighted graphs can also be used to describe the similarity relationships between ”point clouds” of vectors. Many approaches for machine learning or pattern recognition problems involve associating each data instance with a collection of feature vectors that hopefully encapsulate sufficient information about the data point to solve the problem at hand.

graph Laplacian : Of key importance for graph theory is the graph Laplacian operator $\mathcal{L}$. The non normalized Laplacian is defined as $\mathcal{L} = D - A$. It can be verified that for any $f \in \mathbb{R}^N$, $\mathcal{L}$ satisfies

$$(\mathcal{L}f)(m) = \sum_{m \backsim n} a_{m,n} \cdot (f(m) - f(n))$$

where the sum over $m \backsim n$ indicates summation over all vertices n that are connected to the vertex m, and $a_{m,n}$ denotes the weight of the edge connecting m and n. For a graph arising from a regular mesh, the graph Laplacian corresponds to the standard stencil approximation of the continuous Laplacian (with a difference in sign). Some authors define and use an alternative, a normalized form of the Laplacian, defined as

$$\mathcal{L}^{norm} = D^{-1/2}\mathcal{L}D^{-1/2} = I - D^{-1/2}AD^{-1/2}$$

It should be noted that $\mathcal{L}$ and $\mathcal{L}^{norm}$ are not similar matrices, in particular, their eigenvectors are different.

5.3 Graph convolutional networks

In this section, we provide theoretical motivation for a specific graph-based neural network model $f(X, A)$ that we will use in the rest. We consider a multi-layer Graph Convolutional Network (GCN) with the following layer-wise propagation rule :

$$H^{(l+1)} = \sigma(\widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}H^{(l)}W^{(l)})$$

Here, $\widetilde{A} = A + I_N$ is the adjacency matrix of the undirected graph G with added self-connections. I_N is the identity matrix, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$ and and $W^{(l)}$ is a layer-specific trainable weight matrix. $\sigma(.)$ denotes an activation function, such as the $ReLU(.) = max(0,.)$. $H^{(l)} \in \mathbb{R}^{N \times D}$ is the matrix of activations in the l^{th} layer ; $H^{(0)} = X$. In the following, we show that the form of this propagation rule can be motivated via a first-order approximation of localized spectral filters on graphsHammond *et al.* (2011); Defferrard *et al.* (2016).

5.3.1 Graph convolution

We consider spectral convolutions on graphs defined as the multiplication of a signal $x \in \mathbb{R}^N$ (a scalar for every node) with a filter $g_\theta = diag(\theta)$ parameterized by $\theta \in \mathbb{R}^N$ in the Fourier domain, i.e. :

$$g_\theta \star x = U g_\theta U^T x, \tag{5.1}$$

where U is the matrix of eigenvectors of the normalized graph Laplacian $\mathcal{L}^{norm} = I - D^{-1/2} A D^{-1/2} = U \Lambda U^T$ with a diagonal matrix of its eigenvalues Λ and $U^T x$ being the graph Fourier transform of x. We can understand g_θ as a function of the eigenvalues of $\mathcal{L}^{norm}$, i.e $g_\theta(\Lambda)$. Evaluating Eq. 5.1 is computationally expensive, as multiplication with the eigenvector matrix U is $\mathcal{O}(N^2)$. Furthermore, computing the eigendecomposition of $\mathcal{L}^{norm}$ in the first place might be prohibitively expensive for large graphs. To circumvent this problem, it was suggested in Hammond *et al.* (2011) that $g_\theta(\Lambda)$ can be well-approximated by a truncated expansion in terms of Chebyshev polynomials $T_k(x)$ up to K^{th} order :

$$g'_\theta(\Lambda) \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{\Lambda})$$

with a rescaled $\tilde{\Lambda} = \frac{2}{\lambda_{max}} \Lambda - I_N$. λ_{max} denotes the largest eigenvalue of $\mathcal{L}^{norm}$. $\theta' \in \mathbb{R}^K$ is now a vector of Chebyshev coefficients. The Chebyshev polynomials are recursively defined as $T_k(x) = 2x T_{k-1}(x) - T_{k-2}(x)$, with $T_0(x) = 1$ and $T_1(x) = x$. The reader is referred to Hammond *et al.* (2011) for an in-depth discussion of this approximation.

Going back to our definition of a convolution of a signal x with a filter g'_θ , we now have :

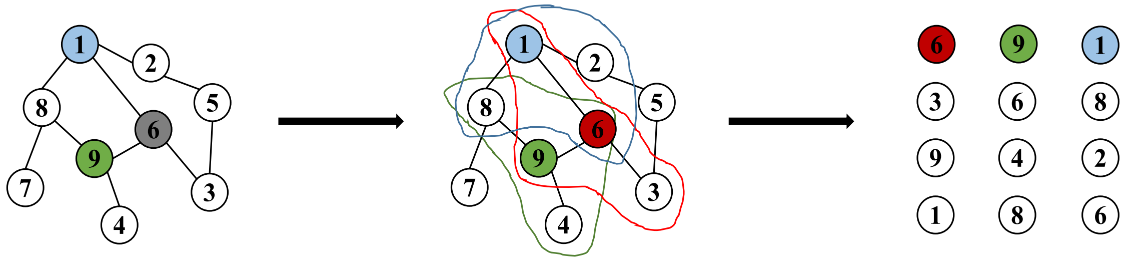
$$g'_\theta \star x \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{L}) x, \tag{5.2}$$

with $\tilde{L} = \frac{2}{\lambda_{max}} L - I_N$; as can easily be verified by noticing that $(U \Lambda U^T)^k = U \Lambda^k U^T$. Note that this expression is now K-localized since it is a K^{th}- order polynomial in the Laplacian, i.e. it depends only on nodes that are at maximum K steps away from the central node (K^{th}-order neighborhood). The complexity of evaluating Eq. 5.2 is $\mathcal{O}(\varepsilon)$, i.e. linear in the number of edges. Defferrard *et al.* (2016) use this K-localized convolution to define a convolutional neural network on graphs.

Convolution is a key process in a graph convolutional network, it allows us to highlight local receptive features. When applying convolution on graphs, two important steps are considered :

— Determine the neighboring nodes around each given graph node for the convolution process : build locally connected subsets of the global graph. These subsets can be obtained using search strategies. We use the breadth-first search to enlarge the neighborhood nodes on the graph, thus, having sets of each node in the graph. Fig. 5.1(a) shows the graph representing the 3D object of and Fig. 5.1(b) shows the neighbourhood sets of each source node.

— Arrange the order of execution of the convolution process on each set : it is not evident to determine the spatial orders of the nodes on the 3D shape. It is thus crucial to reasonably sort the nodes in the subsets to ensure that the elements can be convolved by the same rules and the convolution can better activate features For each node. To do so, the L2 similarity method is used. Fig. 5.1(b) shows the final convolution order for each subset.



(a) 3D shape presented by a graph

(b) Subgraphs: sets of nodes neighborhood

(c) Convolution order

FIGURE 5.1 – Convolution on graph. (a) The 3D object represented by a graph ; (b) subsets searched by a breadth-first search (c) Convolution order of each subset.

5.3.2 Multiple convolution layers model

A neural network model based on graph convolutions can, therefore, be built by stacking multiple convolutional layers of the form of Eq. 5.2, each layer followed by a point-wise non-linearity. Now, imagine we limited the layer-wise convolution operation to K = 1 (see Eq. 5.2), i.e. a function that is linear w.r.t. $\mathcal{L}^{norm}$ and therefore a linear function on the graph Laplacian spectrum.

In this way, we can still recover a rich class of convolutional filter functions by stacking multiple such layers, but we are not limited to the explicit parameterization given by, e.g., the Chebyshev polynomials. We intuitively expect that such a model can alleviate the problem of overfitting in local neighborhood structures for graphs with very wide node degree distributions, such as social networks, citation networks, knowledge graphs, and many other real-world graph datasets. Additionally, for a fixed computational budget, this layer-wise linear formulation allows us to build deeper models, a practice that is known to improve modeling capacity on several domains He *et al.* (2016).

In this linear formulation of a GCN we further approximate $\lambda_{max} \approx 2$, as we can expect that neural network parameters will adapt to this change in scale during training. Under these approximations Eq. 5.2 simplifies to :

$$g'_\theta \star x \approx \theta'_0 x + \theta'_1 (L - I_N) x = \theta'_0 x - \theta'_1 D^{-1/2} A D^{-1/2} x,$$

with two free parameters $theta'_0$ and $theta'_1$. The filter parameters can be shared over the whole graph. Successive application of filters of this form then effectively convolve the k^{th}-order neighborhood of a node, where k is the number of successive filtering operations or convolutional layers in the neural network model.
In practice, it can be beneficial to constrain the number of parameters further to address overfitting and to minimize the number of operations (such as matrix multiplications) per layer. This leaves us with the following expression :

$$g_\theta \star x \approx \theta(I_N + D^{-1/2} A D^{-1/2}) x,$$

with a single parameter $\theta = \theta'_0 = -\theta'_1$. Note that $I_N + D^{-1/2} A D^{-1/2}$ now has eigenvalues in the range $[0, 2]$. Repeated application of this operator can therefore lead to numerical instabilities and exploding/vanishing gradients when used in a deep neural network model. To alleviate this problem, the following renormalization is introduced : $I_N + D^{-1/2} A D^{-1/2} \rightarrow \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2}$, with $\tilde{A} = A + I_N$ and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$.

We can generalize this definition to a signal $X \in \mathbb{R}^{N \times C}$ with C input channels (i.e. a C-dimensional feature vector for every node) and F filters or feature maps as follows :

$$Z = \widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} X \Theta,$$

where $\Theta \in \mathbb{R}^{C \times F}$ is now a matrix of filter parameters and $Z \in \mathbb{R}^{N \times F}$ is the convolved signal matrix. This filtering operation has complexity $\mathcal{O}(|\varepsilon| FC)$, as $\tilde{A}X$ can be efficiently implemented as a product of a sparse matrix with a dense matrix.

5.3.3 Graph Coarsening and pooling

The pooling operation requires meaningful neighborhoods on graphs, where similar vertices are clustered together. Doing this for multiple layers is equivalent to a multi-scale clustering of the graph that preserves local geometric structures. It is however known that graph clustering is NP-hard Bui et Jones (1992) and that approximations must be used. While there exist many clustering techniques, e.g. the popular spectral clustering Von Luxburg (2007), we are most interested in multilevel clustering algorithms where each level produces a coarser graph which corresponds to the data domain seen at a different resolution. Moreover, clustering techniques that reduce the size of the graph by a factor two at each level offers a precise control on the coarsening and pooling size. In this work, we make use of the coarsening phase of the Graclus multilevel clustering algorithm Dhillon *et al.* (2007), which has been shown to be extremely efficient at clustering a large variety of graphs. To compute coarser scales of a given graph, Graclus uses a greedy algorithm. It picks an unmarked vertex i and match it with one of its unmarked neighbors j. The matched vertices are then marked and the coarsened weights are set as the sum of their weights. This process is repeated untill all nodes are explored. The coarsing divides the number of nodes by approximately two.

Pooling operations are carried out many times and must be efficient. After coarsening, the vertices of the input graph and its coarsened versions are not arranged in any meaningful way. Hence, a direct application of the pooling operation would need a table to store all matched vertices. That would result in a memory inefficient, slow, and hardly parallelizable implementation. It is however possible to arrange the vertices such that a graph pooling operation becomes as efficient as a $1D$ pooling according to two steps : create a balanced binary tree, and rearrange the vertices. After coarsening, each node has either two children, if it was matched at the finer level, or one, if it was not, i.e the node was a singleton. From the coarsest to the finest level, fake nodes, i.e. disconnected nodes, are added to pair with the singletons such that each node has two children. This structure is a balanced binary tree : regular nodes (and singletons) either have two regular nodes (e.g. level 1 vertex 0 or one singleton and a fake node as children, and fake nodes always have two fake nodes as children. Input signals are initialized with a neutral value at the fake nodes, e.g. 0 when using a ReLU activation with max pooling. Because these nodes are disconnected, filtering does not impact the initial neutral value. While those fake nodes do artificially increase the dimensionality thus the computational cost, we found that, in practice, the number of singletons left by Graclus is quite low. Arbitrarily ordering the nodes at the coarsest level, then propagating this ordering to the finest levels, i.e. node k has nodes $2k$ and $2k + 1$ as children, produces a regular ordering in the finest level. Regular in the sense that adjacent nodes are hierarchically merged at coarser levels. Pooling such a rearranged graph signal is analog to pooling a regular 1D signal. Fig. 5.2 shows an example of the process.

5.3.4 Semi supervised node classification

A large number of approaches for semi-supervised learning using graph representations have been proposed in recent years, most of which fall into two broad categories : methods that use some form of explicit graph Laplacian regularization and graph embedding-based approaches. Prominent examples for graph Laplacian regularization include label propagation Zhu *et al.* (2003), manifold regularization Belkin *et al.* (2006) and deep semi-supervised embedding Weston *et al.* (2012).
Recently, attention has shifted to models that learn graph embeddings with methods inspired by the skip-gram model Mikolov *et al.* (2013). DeepWalk Perozzi *et al.* (2014) learns embeddings via the prediction of the local neighborhood of nodes, sampled from random walks on the graph. LINE Tang *et al.* (2015) and node2vec Grover et Leskovec (2016) extend DeepWalk with more sophisticated random walk or breadth-first search schemes. For all these methods, however, a multistep pipeline including random walk generation and semi-supervised training is required where each step has to be optimized separately. Planetoid Yang *et al.* (2016) alleviates this by injecting label information in the process of learning embeddings.

Neural networks that operate on graphs have previously been introduced in Gori *et al.*

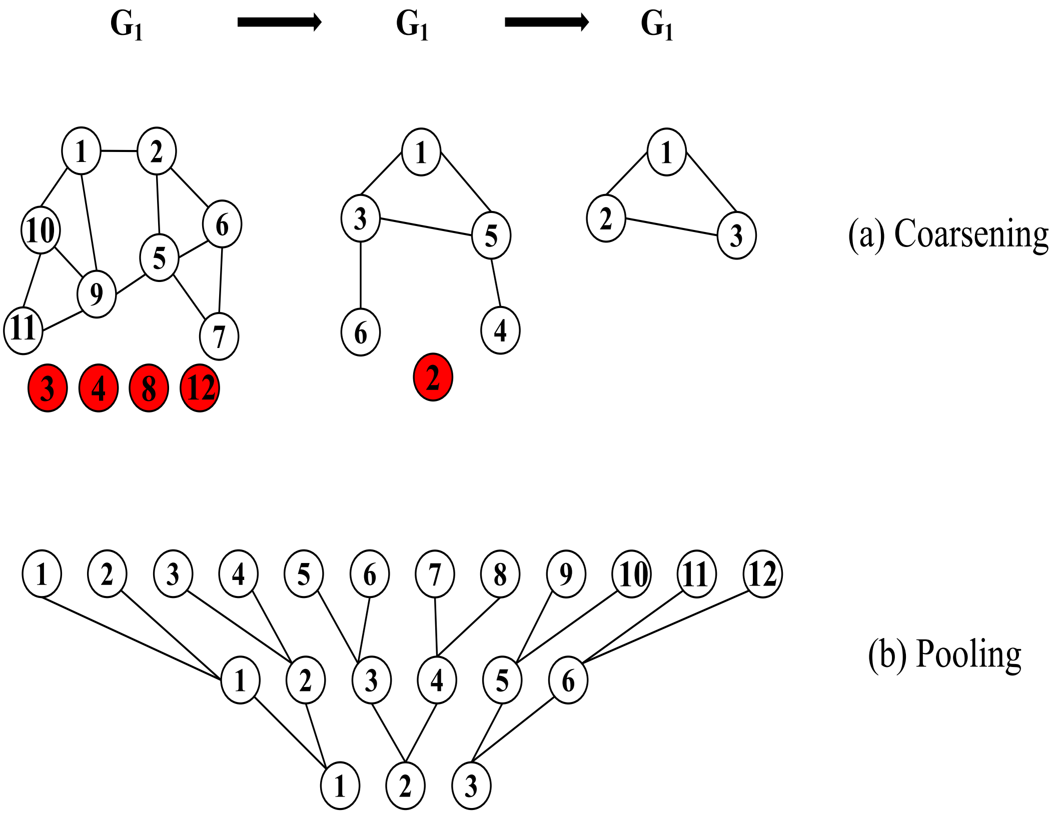


FIGURE 5.2 – Example of Graph Coarsening and Pooling

(2005) ; Scarselli *et al.* (2008) as a form of recurrent neural network. Their framework requires the repeated application of contraction maps as propagation functions until node representations reach a stable fixed point. This restriction was later alleviated in Li *et al.* (2015) by introducing modern practices for recurrent neural network training to the original graph neural network framework. Duvenaud *et al.* (2015) introduced a convolution-like propagation rule on graphs and methods for graph-level classification. Their approach requires learning node degree-specific weight matrices which do not scale to large graphs with wide node degree distributions. Our model instead uses a single weight matrix per layer and deals with varying node degrees through an appropriate normalization of the adjacency matrix.

Having introduced a simple, yet flexible model $f(X, A)$ for efficient information propagation on graphs, we can return to the problem of semi-supervised node classification. As outlined in the introduction, we can relax certain assumptions typically made in graph-based semi-supervised learning by conditioning our model $f(X, A)$ both on the data X and on the adjacency matrix A of the underlying graph structure. We expect this set to be especially powerful in scenarios where the adjacency matrix contains information not present in the data X, such as citation links between documents in a citation network or relations in a knowledge graph.

In the following, we consider a two-layer GCN for semi-supervised node classification on a graph with a symmetric adjacency matrix A (binary or weighted). We first calculate

$\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ in a pre-processing step. Our forward model then takes the simple form :

$$Z = f(X, A) = \text{softmax}\left(\hat{A}ReLu\left(\hat{A}XW^0\right)W^1\right)$$

Here, $W^0 \in \mathbb{R}^{C\times H}$ is an input-to-hidden weight matrix for a hidden layer with H feature maps. $W^1 \in \mathbb{R}^{H\times F}$ is a hidden-to-output weight matrix. The softmax activation function, defined as :

$$\text{softmax}(x_i) = \frac{1}{z}\exp(x_i)$$

with $z = \sum_i \exp(x_i)$ is applied row-wise. For semi-supervised multiclass classification, we then evaluate the cross-entropy error over all labeled examples :

$$L = \sum_{l\in Y_L}\sum_{f=1}^{F} Y_{lf}\ln Z_{lf},$$

where Y_L is the set of node indices that have labels.
The neural network weights $W_{(0)}$ and $W_{(1)}$ are trained using gradient descent.

5.4 No-reference quality assessment using graph convolutional neural network

In this section, we describe the proposed method for mesh visual quality assessment based on graph convolutional network. The first step of our method consists of the construction of the graph from the 3D mesh. To do so, we present the graph with an adjacency matrix that indicates whether pairs of nodes are connected or not. After that, we use a graph convolutional network for mesh quality assessment. In this work, the network consists of three modules :

- The convolution module that performs the convolution operation over the input graph.
- The pooling module that is used to provide the final feature representation by aggregating the convolution vectors.
- The classification module that estimates the perceived visual quality based on the level of deformation.

We present in Fig. 5.3 the block diagram of the proposed method depicting the different modules.

5.4.1 Graph network representation and adjacency matrix

The 3D object is transformed into a graph network. This latter is presented by an adjacency matrix A of size $n \times n$ where n is the number of nodes. If two vertices v_i and v_j are adjacent (connected) the value of $A_{i,j}$ is set to 1, otherwise, the value is set to 0. Fig. 5.4 illustrates an example of a graph network representation of a 3D mesh.

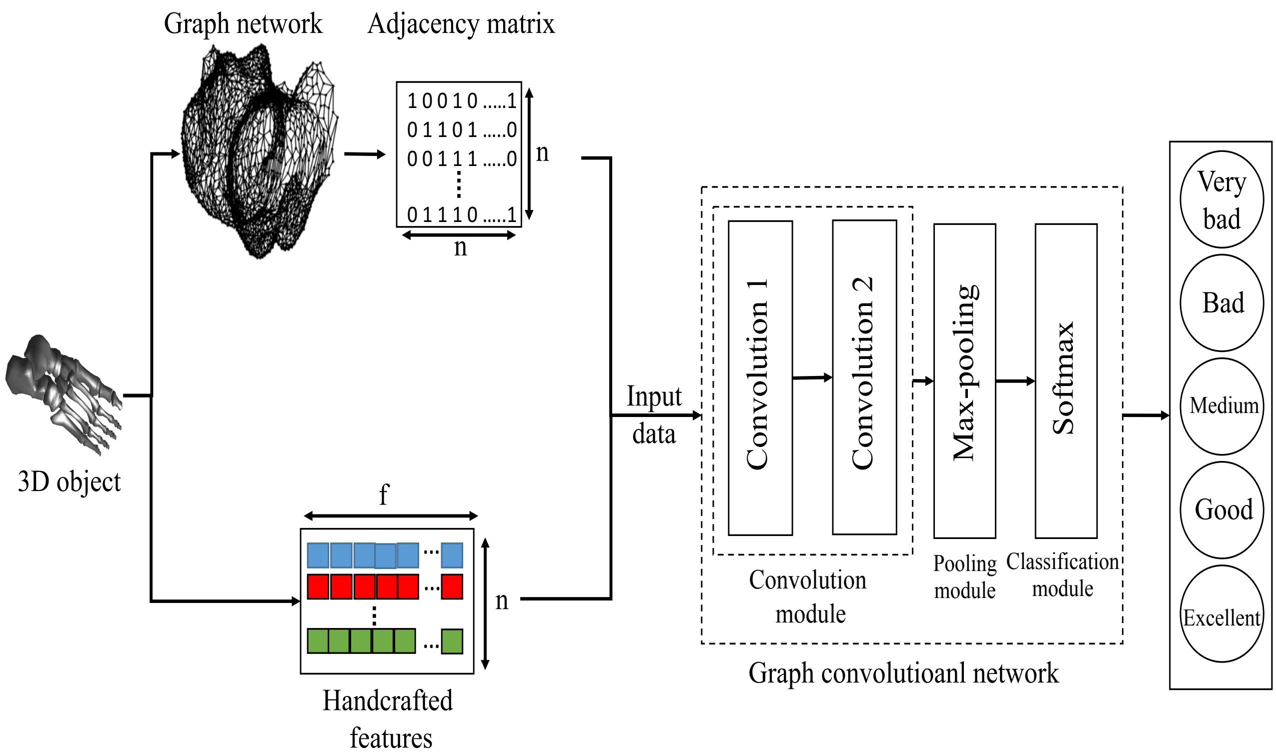
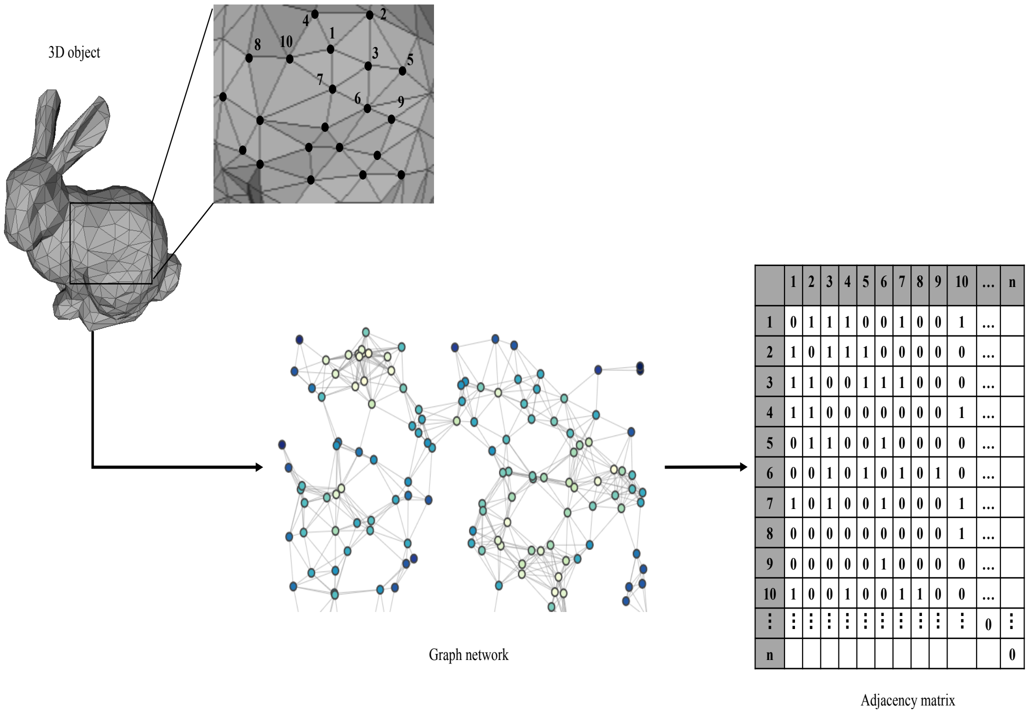


Figure 5.3 – The overall scheme of the proposed method.



	1	2	3	4	5	6	7	8	9	10	...	n
1	0	1	1	1	0	0	1	0	0	1	...	
2	1	0	1	1	1	0	0	0	0	0	...	
3	1	1	0	0	1	1	1	0	0	0	...	
4	1	1	0	0	0	0	0	0	0	1	...	
5	0	1	1	0	0	1	0	0	0	0	...	
6	0	0	1	0	1	0	1	0	1	0	...	
7	1	0	1	0	0	1	0	0	0	1	...	
8	0	0	0	0	0	0	0	0	0	1	...	
9	0	0	0	0	0	1	0	0	0	0	...	
10	1	0	0	1	0	0	1	1	0	0	...	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	0	⋮
n												0

Figure 5.4 – Graph representation and adjacency matrix from a 3D object.

5.4.2 Handcrafted features

In our work, we use four geometric features : curvature, angles, Laplacian of Gaussian curvature, and saliency.

— **Curvature** describes the deviation amount of the surface of being flat. In this work, we use the maximum curvature amplitude C_{max} and the minimum curvature amplitude C_{min}. The mean curvature $C_{mean} = (C_{max} + C_{min})/2$ as well as the Gaussian curvature as $C_{gauss} = C_{max} \times C_{min}$ are also computed. The curvature features are labeled as C.

— **Laplacian of Gaussian curvature** corresponds to the Laplacian operator applied to the Gaussian curvature field. This feature is labeled here as LoG.

— **Angle** is computed for each edge and corresponds to the angle between the normals of its adjacent faces. It represents the structural aspect of the mesh. The computed angle values are averaged to obtain a scalar value for each vertex. This feature is labeled here as An.

— **Saliency** is a perceptual concept that describes the attention of our HVS to some regions due to its specificities (curvature, orientation, and so on). This feature is labeled here as S.

Fig. 5.5 presents a graphic illustration of the used features.

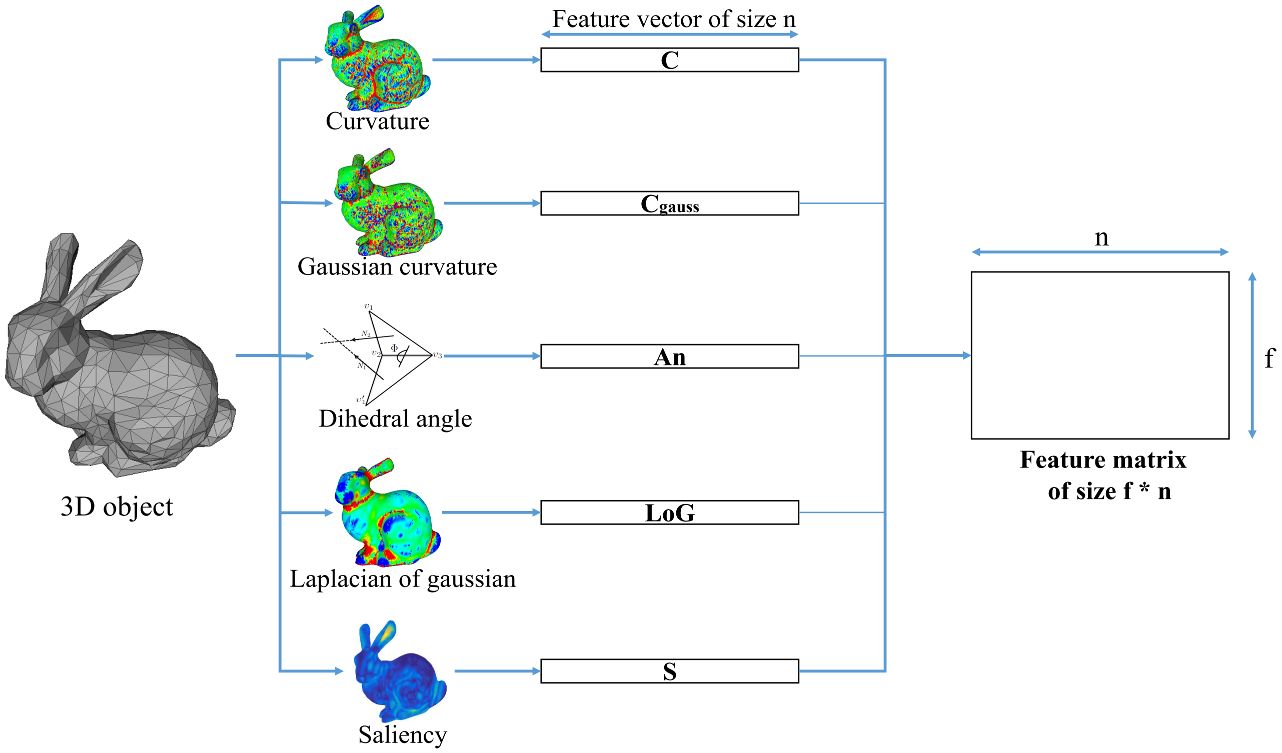


FIGURE 5.5 – Handcrafted features preparation.

5.4.3 Graph convolution

Let $G = \{V, \mathcal{E}\}$ denotes the graph representing the distorted mesh with $v_i \in V$ and $(v_i, v_j) \in \mathcal{E}$ are the sets of nodes and edges of G respectively. $A \in \mathbb{R}^{N \times N}$ denotes the adjacency matrix of G and $D_{ii} = \sum_j A_{ij}$ is the degree matrix. The identity matrix is denoted by I_N.

The graph convolution is defined as the process of filtering a signal $x \in \mathbb{R}^N$ with a filter f_θ.

$$Y = f_\theta * x = \theta(I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}})x \tag{5.3}$$

The operator $I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ may cause numerical stabilities when used in a deep network module. To prevent this issue the following normalization is conducted : $I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \to \widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}}$. Where $\widetilde{A} = A + I_N$ and $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$.

Finally, the generalized graph convolution is defined for a signal $X \in \mathbb{R}^{N \times C}$ with C input channels representing C-dimensional feature vector for every node and F filters as follows :

$$Y = \widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} X \Theta \tag{5.4}$$

where $\Theta \in \mathbb{R}^{C \times F}$ is a matrix of filter parameters and $Y \in \mathbb{R}^{N \times F}$ is the convolved matrix.

5.4.4 Pooling

The graph convolutional network can be composed of K convolutional layers that provide a rich feature vectors representation.

The next step of our method is to summarize the feature maps obtained by the convolution into a single vector representation. To do so, the feature maps are fed into a pooling module to aggregate such convolution vectors.

In our work, we perform a max-pooling which is computed by selecting the max value over the entire convolution vector sequence.

5.4.5 Training and classification

To estimate the visual quality of a distorted mesh we rely on the problem of node classification. In our work, five classes are considered depending on the given mean opinion score for each database : very bad, bad, medium, good and excellent quality. The quality of each node is estimated and the overall quality of the whole distorted mesh is obtained by considering the class with the maximum number of nodes.

The training process is conducted using the leave one out cross-validation. For a given database, a training model is built using all the existing objects in the repository except one object and its corresponding distortions.

The negative log-likelihood is minimized over the training set using the stochastic gradient

descent (SGD) and backpropagation is used to compute the gradients while dropout is employed to avoid over-fitting.
For the test phase, the excluded objects serve as a test set, and the trained network classifies the given distorted meshes concerning their visual quality. The obtained classes are then compared with the subjective labels using correlation measurements.

5.5 Results and discussion

5.5.1 Effect of the geometric attributes

In this section, evaluate the impact of the extracted handcrafted feature on the performance. We tested several sets used as input to the GCN. Table 5.1 shows the results obtained. We stated previously that we used in our method four geometric attributes to construct the feature matrix. We tested more attributes before we arrive at the best combination. In this section we provide the obtained results from different features combinations (Table. 5.1). The tested features are minimum and maximum directions (Umin and Umax), curvature-based (Cmin, Cmax, Cmean, and Cgauss), Normal (Nor), Saliency (S), Angle (An), and Laplacian of Gaussian (LoG).

Tableau 5.1 – Correlation coefficients r_s (%) and r_p (%) using different feature combinations on the general-purpose database.

Features	Armadillo		Dyno		Venus		Rocker		All models	
	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
C	75.6	78.1	82.1	81.8	88.1	86.0	74.7	73.1	76.5	74.5
C + An	86.3	85.9	88.0	88.6	90.7	91.2	86.9	87.8	83.7	82.9
C + An+ S	90.9	89.1	90.1	88.3	92.7	89.8	87.1	86.0	86.7	85.7
C + An+ LoG	93.8	92.7	88.9	90.5	91.3	91.0	90.7	88.8	87.9	88.6
C + An+ LoG + S	92.4	93.2	90.5	91.0	94.2	93.6	89.5	89.2	**90.3**	**89.8**
C + An+ LoG + S + U	90.5	91.4	92.5	92.3	91.4	91.2	88.9	88.4	88.5	87.3
C + An+ LoG + S + U + Nor	91.2	90.1	89.1	88.9	92.8	93.1	88.1	87.0	87.9	86.3

As can be seen, the performance using curvature-based and angle feature obtained good correlations. Adding the Saliency considerably improves the performance. Similar results are obtained when LoG is added. Unfortunately, adding the directions (Umin and Umax) and the normal (Nor) does not improve the results. The best performance is obtained using the four features (C + Ang + LoG + S) leads to the best correlation scores. Therefore, we only rely on Cmin, Cmax, Cmean, Cgauss, Ang, LoG, and S to construct the matrix of features.

5.5.2 Comparison

To evaluate the performance of our method, a comparative analysis is conducted. Our method is compared to the state of the art methods including full reference, reduced reference, and no reference methods :

The correlation coefficients values r_s and r_p on the LIRIS masking, LIRIS/EPFL General-purpose, UWB compression and the IEETA simplification databases are listed respectively in Tables. 5.2- 5.3.

Tableau 5.2 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the LIRIS/EPFL general-purpose database.

Type	Metric	Armadillo		Dyno		Venus		Rocker		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	69.5	30.2	30.9	22.6	1.6	0.8	18.1	5.5	13.8	1.3
	RMS Cignoni *et al.* (1998)	62.7	32.3	0.3	0.0	90.1	77.3	7.3	3.0	26.8	7.9
	MSDM2 Lavoué (2011)	81.6	85.3	85.9.4	85.7	89.3	87.5	89.6	87.2	80.4	81.4
	TPDM Torkhani *et al.* (2014)	84.5	78.8	92.2	89.0	90.6	91.0	92.2	91.4	**89.6**	**89.2**
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	65.8	35.7	62.7	35.7	71.6	46.6	87.5	53.2	69.3	38.4
	3DWPM2 Corsini *et al.* (2007)	74.1	43.1	52.4	19.9	34.8	16.4	37.8	29.9	49.0	24.6
	FMPD Wang *et al.* (2012)	75.4	83.2	89.6	88.9	87.5	83.9	88.8	84.7	**81.9**	**83.5**
No-Reference	NR-SVR Abouelaziz *et al.* (2016b)	76.8	91.5	78.6	84.1	85.7	88.6	86.2	86.6	81.5	87.8
	NR-GRNN Abouelaziz *et al.* (2016a)	87.1	**97.3**	91.2	**94.1**	86.3	85.0	78.6	74.8	86.2	88.7
	NR-CNN1 Abouelaziz *et al.* (2017)	87.2	84.3	86.4	86.2	92.2	85.6	91.3	85.2	83.6	82.7
	NR-CNN2 Abouelaziz *et al.* (2018)	93.4	95.6	86.2	84.3	**94.1**	90.3	80.4	82.2	81.8	82.5
	BMQI Nouri *et al.* (2017)	20.1	-	83.5	-	88.9	-	92.7	-	78.1	-
	CNN-BMQA	90.6	92.4	88.3	86.3	93.1	92.6	88.9	89.0	**90.0**	**92.0**
	CNNs-CMP	93.4	92.9	91.6	90.9	88.9	89.4	92.6	93.9	**92.6**	**91.3**
	SCNNs-CMP	**95.8**	95.6	**93.6**	92.9	93.4	**91.3**	**94.5**	**95.2**	94.4	94.8
	MVQ-GCN	91.8	92.5	87.7	84.5	93.7	91.9	89.6	88.4	**89.3**	**88.6**

From Tables. 5.2- 5.3. we can notice that measures HD and RMS metrics perform the worst since they rely only on geometric distances without taking into consideration the main operations of the HVS. contrariwise MSDM2 and TPDM achieve better prediction compared to the geometric measures as proven by the high correlation coefficients. These methods incorporate the perceptual information, thus, the perceptual information is included. FMPD provides the highest scores concerning RR methods. This method includes a roughness measure which is an important perceptual feature in mesh processing.

Our method shows excellent performance on the two subjectively-rated MVQ databases, as proven by its high scores on the individual models as well as on the whole repositories.

On the LIRIS masking database, our method provides the highest Pearson correlation coefficients on the whole corpus (r_p = 90.9%) and the second-highest Spearman score (r_s = 91.7%) with a slight difference compared to NR-CNN2 (r_s = 92.0%). Besides, the performance of our method outperforms the most effective FR and RR methods.

The General-purpose database is the largest MVQ database so far, this database contains the highest number of distorted models (21 distorted version for each model as well as a variety of distortion types). In this database, our method shows good performance and provide competitive correlation scores (r_s = 89.3% and r_p = 88.6%) that contend many effective FR, RR, and NR methods. The high correlation scores provided by our

Tableau 5.3 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the LIRIS masking database.

Type	Metric	Armadillo		Lion		Bimba		Dyno		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
Full Reference	HD Aspert *et al.* (2002)	48.6	37.7	71.4	25.1	25.7	7.5	48.6	31.1	26.6	4.1
	RMS Cignoni *et al.* (1998)	65.6	44.6	71.4	23.8	71.4	21.8	71.4	50.3	48.8	17.0
	MSDM2 Lavoué (2011)	81.1	88.6	93.5	94.3	96.8	100	95.6	100	87.3	89.6
	TPDM Torkhani *et al.* (2014)	91.4	88.6	88.4	82.9	97.1	100	71.1	100	**88.6**	**90.0**
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	58.0	41.8	20.0	9.7	20.0	8.4	66.7	45.3	29.4	10.2
	3DWPM2 Corsini *et al.* (2007)	48.6	37.9	38.3	22.0	37.1	14.4	71.4	50.1	37.4	18.2
	FMPD Wang *et al.* (2012)	94.2	88.6	93.5	94.3	98.9	100	96.9	94.3	**80.8**	**80.2**
No-Reference	NR-SVR Abouelaziz *et al.* (2016b)	89.5	84.7	**100**	96.3	94.2	93.6	94.4	89.7	90.4	91.2
	NR-GRNN Abouelaziz *et al.* (2016a)	82.3	80.5	94.1	97.0	90.2	94.3	78.2	82.3	90.2	82.4
	NR-CNN1 Abouelaziz *et al.* (2017)	95.2	97.6	89.4	91.6	93.4	98.7	96.3	89.9	88.2	85.4
	BMQI Nouri *et al.* (2017)	94.3	-	94.3	-	**100**	-	83.0	-	78.1	-
	CNN-BMQA	92.5	93.1	90.6	89.6	88.5	90.2	92.3	90.6	**91.4**	**90.8**
	CNNs-CMP	94.0	94.6	92.6	91.6	90.7	91.2	93.4	92.6	**93.2**	**92.8**
	SCNN-BMQA	92.4	91.7	92.2	93.1	97.9	97.3	93.4	92.6	95.5	94.3
	SCNNs-CMP	96.2	95.5	93.1	92.4	92.5	92.8	94.2	94.0	95.8	95.9
	MVQ-GCN	92.6	91.2	93.5	89.9	88.6	89.8	89.3	88.7	**91.7**	**90.9**

method in this database appear to be a good indicator of the forcefulness of our method in MVQ assessment.

Tableau 5.4 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the UWB compression database.

Type	Metric	Bunny		James		Jessy		Nissan		Helix		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p		
Full Reference	HD Aspert *et al.* (2002)	34.1	52.2	-16.8	6.8	-23.6	12.5	14.4	23.6	45.1	46.4	10.6	28.3
	RMS Cignoni *et al.* (1998)	34.2	20.9	14.0	10.8	0.0	14.8	17.8	29.7	46.9	44.6	22.0	24.1
	MSDM2 Lavoué (2011)	97.4	90.1	82.6	69.2	84.3	63.1	84.4	73.1	98.1	94.7	89.3	78.0
	TPDM Torkhani *et al.* (2014)	95.1	96.5	90.8	73.6	85.8	75.8	82.7	73.4	98.7	95.0	**91.5**	**82.9**
Reduced Reference	3DWPM1 Corsini *et al.* (2007)	94.7	93.4	77.3	72.3	87.2	89.5	63.6	59.3	98.0	95.2	84.1	81.9
	3DWPM2 Corsini *et al.* (2007)	96.0	91.2	76.9	65.3	86.9	85.9	56.3	67.6	95.5	94.3	82.3	80.9
	FMPD Wang *et al.* (2012)	94.2	89.6	95.3	91.2	63.3	60.0	92.4	77.5	98.4	90.8	88.8	81.8
	DAME Váša et Rus (2012)	96.8	93.4	95.7	93.4	84.4	70.5	93.9	75.3	96.6	95.2	**93.5**	**85.6**
No-Reference	CNN-BMQA	92.3	91.9	88.9	89.3	91.3	92.4	90.1	90.8	91.6	89.5	**90.1**	**88.3**
	CNNs-CMP	91.3	90.6	93.2	92.6	89.3	88.6	90.1	90.3	91.6	89.6	90.2	90.9
	SCNN-BMQA	95.8	91.7	96.2	95.6	92.3	90.5	88.7	84.7	96.7	94.6	90.4	88.2
	SCNNs-CMP	95.6	94.8	**92.5**	90.6	92.5	87.1	88.7	**89.0**	90.7	90.4	92.7	93.8
	MVQ-GCN	93.4	92.3	90.6	90.9	91.2	91.0	88.6	89.5	90.1	91.2	**90.5**	**87.7**

Tableau 5.5 – Correlation coefficients r_s (%) and r_p (%) of different objective metrics on the IEETA simplification database.

Type	Metric	Bones		Bunny		Head		Lung		Strange		All models	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p		
Full Reference	HD Aspert *et al.* (2002)	92.0	94.3	37.8	39.5	72.8	88.6	80.6	88.6	52.3	37.1	50.5	49.4
	RMS Cignoni *et al.* (1998)	86.4	94.3	94.5	77.1	49.6	42.9	89.0	100	90.4	88.6	59.6	70.2
	MSDM2 Lavoué (2011)	98.3	94.3	98.1	77.1	88.9	88.6	92.3	60.0	99.0	94.3	**89.2**	86.7
	TPDM Torkhani *et al.* (2014)	99.0	94.3	98.0	94.3	63.1	65.7	98.6	94.3	98.7	94.3	86.9	**88.2**
Reduced Reference	FMPD Wang *et al.* (2012)	96.0	88.6	98.0	94.3	70.4	65.7	95.5	88.6	96.0	65.7	89.3	87.2
No-Reference	CNN-BMQA	93.2	92.4	90.3	91.0	88.6	89.3	89.0	88.9	91.5	92.1	**90.4**	**90.2**
	CNNs-CMP	92.9	91.9	93.1	93.0	89.6	88.6	92.0	91.6	94.3	93.2	**90.6**	**90.4**
	SCNN-BMQA	96.8	93.7	96.7	90.9	96.4	93.6	94.3	89.6	95.4	93.5	90.4	90.5
	SCNNs-CMP	91.3	88.9	91.1	92.4	**91.8**	**91.5**	95.3	89.4	91.1	88.9	91.0	91.1
	MVQ-GCN	90.4	89.6	91.0	91.8	90.6	92.1	93.6	92.5	89.1	88.8	**89.9**	**89.4**

5.5.3 Cross dataset evaluation

In this section, we investigate the generalization ability of our method to predict the quality. To do so, we perform a cross-database evaluation by training our network on the General-purpose database and using the other databases for the test. We choose this database for the training process because it contains the highest number of distorted models and a rich variety of distortion types. Table. 5.6 shows the results of our evaluation. This table presents the correlation coefficients of each 3D object in the three tested databases (i.e LIRIS masking, UWB compression, and IEETA simplification) as well as the scores for the whole repositories.

Tableau 5.6 – Correlation coefficients r_s (%) and r_p (%) obtained by training on the General-purpose and testing on LIRIS masking, UWB compression and IEETA simplification.

Databases	Masking	Compression	Simplification
r_s	90.8	87.2	84.8
r_p	91.2	86.3	83.3

As we can see, our network successfully estimates the perceived visual quality using the cross dataset evaluation as proven by the high correlation coefficients obtained. These results ensure the generalization ability of our method.

5.6 Conclusion

We developed in this chapter a model-based method to accurately estimate the perceived visual quality of distorted meshes. This approach takes advantage of the operation of convolution using directly the 3D mesh itself without rendering 2D views. We relied on a graph convolutional network. The graph consists of two convolutional layers, a pooling layer, and a classification layer. The network is fed by a graph represented by an adjacency matrix and a feature matrix containing four geometric features (curvature, angles, Laplacian of Gaussian curvature, and Saliency). We tested several features before arriving at the best features configuration. The proposed method successfully predicts the visual quality of distorted meshes as proven by the high correlations with human judgment.

CONCLUSION AND FUTUR WORK

Conclusion

Recently, 3D models have been widely used in several applications, and the interest in perceptual visual quality assessment of 3D meshes has become an important issue. The majority of existing methods are whether full-reference or reduced-reference. In other words, the visual quality in these methods is evaluated based on the reference mesh and its corresponding distorted version by computing geometric distances or comparing some extracted features including perceptual information. These methods require inevitably the presence of the reference mesh which is not available in most practical situations. However, the development of a no-reference MVQ assessment method becomes crucial to remedy this problem. Our study focused on no-reference quality assessment by developing objective and effective metrics. The objective of developing such metrics is to provide algorithms capable of blindly predicting the quality perceived by a human observer.

We developed two methods based on hand-crafted features extracted from the distorted meshes and machine learning techniques to predict the visual quality. The first method so-called NR-SVR is based on dihedral angles features and the support vector regression (SVR). The second method so-called NR-GRNN based on mean curvature features and the general regression neural network (GRNN). We take into consideration the main function of the HVS by including a visual masking module.

We are also interested in deep learning approaches for feature learning and quality estimation. To this end, several CNNs architectures are used. First, a network from scratch is adopted with several parameters variety to select the best network combination. We called this method CNN-BMQA. Second, We fused automatically learned features extracted from fine-tuned networks, and the compact multilinear pooling is used to combine the vectors to be used for the quality prediction. We called this method CNN-CMP.

To better estimate the perceived visual quality, it is crucial to include more perceptual information to take into account the main operations of the human visual system. Visual saliency is a perceptual aspect that describes the attention of our HVS. We investigated in our study the usage of CNN architectures with 3D visual saliency to estimate the perceived visual quality of distorted meshes. We used a saliency-based patch selection strategy the select only the relevant information of the meshes for the training process. We adopted the network architectures used before and we called the new methods SCNN-BMQA for the network from scratch and CNN-CMP for the fine-tuned networks and CMP combination.

NR-SVR

We have proposed an efficient and effective method for a no-reference mesh quality assessment. Given only a distorted mesh, the proposed scheme extracts dihedral angles as relevant information that describes the structural information. The extracted feature vector is then modulated with a visual masking function which is an important characteristic of the human visual system. The obtained vector is then modeled by three statistical distributions : Gamma, Weibull, and Rayleigh in order to construct three sets of feature vectors with only statistical parameters instead of using whole features values. This step is very crucial to reduce computational time. Once the feature vectors are constructed, the proposed scheme predicts three intermediate scores using the support vector regression (SVR) according to the three distributions. Finally, the overall objective quality score is computed by a weighted sum of the intermediate scores.

NR-GRNN

We proposed another no reference method based on machine learning and handcrafted features. We used the general regression neural network (GRNN) fed by the mean curvature, which considered as a perceptual relevant feature representing the visual aspect of a 3D mesh. The network consists of four layers : the first layer is the input layer fed by the discrete curvature extracted from the distorted mesh. The second layer is the pattern layer in which each unit represents a training pattern. The third layer is the summation layer with two outputs. The output layer is the last in the general regression neural network architecture, it computes the quotient of the two outputs of the summation layer to predict the final mesh visual quality score. The network weights are crucial parameters in the architecture. The training process is conducted by the leave-one-out cross-validation and the obtained scores are then compared with the subjective scores using correlation measurements. This method obtained good scores in terms of the correlation with the mean opinion scores provided subjectively, and generally comparable with competitive full reference and reduced reference metrics.

CNN-BMQA and CNNs-CMP

In CNN-BMQA, a CNN architecture is used to learn sets of 2D patches rendered from the 3D mesh. 2D projections are obtained to represent the 3D mesh from multiple views. To do so, virtual cameras are fixed at different angles around the 3D mesh according to the axes X and Y. Then the obtained views are split into 2D small patches of a fixed size. Finally, CNN is used with a regression method to estimate the objective score for each selected patch. The final quality score for the 3D object is obtained by averaging the scores. many network configurations are testes (the number and the size of kernels, and the size of the input data). It is demonstrated from the experiments that the CNN parameters significantly affect the performance of the network. In CNN-CMP three pre-trained deep convolutional neural networks are employed for feature learning : VGG, AlexNet, and

ResNet. Each network is fine-tuned and produces a feature vector. The Compact Multi-linear Pooling (CMP) is used afterward to fuse the retrieved vectors into a global feature representation. Finally, fully connected layers followed by a regression module are used to estimate the quality score. Several tests have been conducted, in particular, different combination strategies are tested and compared, and we show that combining multiple DCNNs increases the performances and we can derive an effective blind MVQ method.

SCNN-BMQA and SCNNs-CMP

To consider more perceptual features, we extended the methods based on CNNs by including 3D visual saliency to the estimation process. We supposed that the human visual system (HVS) is more sensitive to distortions in salient regions, whereas in non-salient regions their influence on the overall judgment can be neglected. In this context, mesh visual saliency is used to indicate the most relevant regions of the 3D mesh. These regions are presented in the form of 2D small patches which are used to feed CNN architectures to learn an effective representation and estimate the perceived visual quality. The methods exploited the sensitivity of the HVS to mesh degradation together with the efficiency of the CNN learning approach. A patch selection technic based on mesh visual saliency is used to select only the relevant patches to feed the CNN architectures. SCNN-BMQA uses the architecture from scratch and SCNNs-CMP uses three pre-trained networks i.e VGG, AlexNet, and ResNet along with The Compact Multi-linear Pooling (CMP) to fuse the retrieved vectors. Several tests have been conducted, in particular, we show that the patch selection strategy is very effective. It is concluded that the distortions in salient regions are more important than in the normal regions, and thus, including the saliency information increases the performance of the quality estimation.

GCN

The perceptual mechanisms on all the previously mentioned methods are applied to the images generated from the 3D data making the evaluation view-dependant also called image-based approach. In this method, we introduced a model-based method using graph convolutional networks (GCN) to work directly on the 3D model itself. Our graph is presented by an adjacency matrix that indicates whether pairs of nodes of the 3D mesh are connected or not. A GCN is used for the learning and quality score estimation. The network consists of the convolution module that performs the convolution operation over the input graph. After that, the pooling module is used to provide the final feature representation by aggregating the convolution vectors. Finally, The classification module is adopted to estimate the perceived visual quality based on the level of deformation. The network is fed by a graph represented by an adjacency matrix and a feature matrix containing four geometric features (curvature, angles, Laplacian of Gaussian curvature, and Saliency). To estimate the visual quality the method relies on the problem of node classification. Five classes are considered depending on the given mean opinion score for each database : very bad, bad, medium, good, and excellent quality. The quality of each node is estimated and the overall quality of the whole distorted mesh is obtained by considering the class with the maximum number of nodes. The proposed method successfully predicts the visual quality of distorted meshes as proven by the high correlations with human judgment.

Perspectives

As perspectives for our research, we distinguish short-term perspectives, such as the study of the effect of rendering parameters and the quality of dynamic meshes. As long-term perspectives, it would be interesting to develop a new mesh quality assessment corpus with an important number of shapes and distortions. Quality of point cloud data is also an interesting field of study in our future work.

Study of the effect of rendering parameters

It is important to incorporate visual rendering parameters such as light in the estimation of the perceived quality of 3D meshes. Such studies will allow us to get closer to the work done on the quality of 2D images and videos. It is also important to conduct researches about the quality of 3D meshes during stereoscopic display. Besides, 3D meshes appear with a texture applied to the surface of the objects and characters presented. It is important to focus on studying the effect of colors and textures on the perception of the quality of 3D meshes.

Quality of dynamic meshes

The area of research of 3D quality assessment has been recently extended to evaluate dynamic meshes (distorted surface animations that are simplified or compressed). As in static meshes, it is crucial to evaluate the perceived visual quality of distorted dynamic meshes. We aim to use deep learning frameworks to extend the proposed methods for static meshes.

development of mesh quality assessment corpus

Several databases have been developed to provide subjective measures of the quality of static meshes. We described these databases in detail in this thesis, they permit the validation of several perceptual metrics for 3D meshes which provide objective measures in correlation with subjective scores. The main drawback of the existing databases is the limited number of objects and distortions. To assess with reliability the performance of objective metrics, it is important to develop a corpus including a large number of distorted meshes with several distortion types annotated subjectively. it would be interesting to present an experimental protocol and describe psycho-visual experiences to develop a strong corpus for mesh visual quality assessment.

Point cloud

As like meshes, 3D data can also be represented with point cloud format. The rapid development and acceptance of various laser scanning technologies have enabled the acquisition of point clouds with important densities. The research of deep learning on 3D point clouds, with an increasing number of methods being proposed to address various problems related to point cloud processing, including 3D shape classification, 3D object detection, and tracking, and 3D point cloud segmentation. As future work, we aim to use deep learning for point cloud data to estimate the perceived visual quality. It would be also useful to use the visual saliency to incorporate the perceptual information.

Scientific production

The works presented in this thesis have been valued in the publications indicated below :

Journal papers (4)

— [J4 (Submitted)] **I. Abouelaziz**, A. Chetouani, M. EL Hassouni, L.J Latecki ,H. Cherifi "Geometric deep learning for mesh visual quality assessment" IEEE Signal Processing Letters (2020)

— [J3] **I. Abouelaziz**, A. Chetouani, M. EL Hassouni, L.J Latecki ,H. Cherifi "No-reference mesh visual quality assessment via ensemble of convolutional neural networks and compact multi-linear pooling" Pattern Recognition (2019)

— [J2] **I. Abouelaziz**, A. Chetouani, M. EL Hassouni, L.J Latecki ,H. Cherifi "3D visual saliency and convolutional neural network for blind mesh quality assessment " Neural Computing and Applications (2019).

— [J1] **I. Abouelaziz**, M. EL Hassouni, H. Cherifi "Blind 3D mesh visual quality assessment using support vector regression" Multimedia Tools and Applications (2018) : 1-22.

National and international conferences (13)

— [C13] **I. Abouelaziz**, A. Chetouani, M. EL Hassouni, L.J Latecki ,H. Cherifi ?Combination Of Handcrafted And Deep Learning-based Features For 3d Mesh Quality Assessment ? in : The 2020 IEEE International Conference on Image Processing, ICIP 2020

— [C12] **I. Abouelaziz**,A. Chetouani, M. EL Hassouni, L.J Latecki ,H. Cherifi "mesh visual quality based on the combination of convolutional neural networks" in : The International Conference on Image Processing Theory, Tools and Applications, (IPTA) 2019.

— [C11] **I. Abouelaziz**,A. Chetouani, M. EL Hassouni ,H. Cherifi " Évaluation de la qualité visuelle des maillages 3D en utilisant un réseau neuronal convolutif et la saillance visuelle " 27ème Colloque francophone de traitement du signal et des images Lille, GRETSI 2019.

— [C10] **I. Abouelaziz**,A. Chetouani, M. EL Hassouni, L.J Latecki ,H. Cherifi " Convolu-tional neural network for blind mesh visual quality assessment using 3D visual saliency " in : The 2018 IEEE International Conference on Image Processing, ICIP 2018.

- [C9] **I. Abouelaziz**, A. Chetouani, M. EL Hassouni, H. Cherifi " A blind mesh visual quality assessment method based on convolutional neural network ", in : IS & T International Symposium on Electronic Imaging, EI 2018 ; Burlingame, California, USA.
- [C8] **I. Abouelaziz**, A. Chetouani, M. EL Hassouni, H. Cherifi " Un réseau neuronal convolutif pour l'évaluation de la qualité des maillages 3D ", in : Compression et Représentation des Signaux Audiovisuels, CORESA 2017 ; Caen, France.
- [C7] **I. Abouelaziz**, A. Chetouani, M. EL Hassouni, H. Cherifi " Mesh visual quality assessment Metrics : A Comparison Study ", in : 13th International Conference on Signal Image Technology and Internet-Based Systems, SITIS 2017 ; Jaipur, India.
- [C6] **I. Abouelaziz**, M. EL Hassouni, H. Cherifi "A convolutional neural network framework for blind mesh visual quality assessment", in : The 2017 IEEE International Conference on Image Processing, ICIP 2017.
- [C5] **I. Abouelaziz**, M. EL Hassouni, H. Cherifi " A Curvature based method for blind mesh visual quality assessment using a general regression neural network", in : 12th International Conference on Signal Image Technology and Internet-Based Systems, SITIS 2016 ; Naples, Italy.
- [C4] **I. Abouelaziz**, M. EL Hassouni, H. Cherifi "No-reference 3D mesh quality assessment based on dihedral angles model and support vector regression", in : International Conference on Image and Signal Processing (ICISP 2016) ; Trois-Rivière - Canada
- [C3] **I. Abouelaziz**, M. Omari, M. EL Hassouni, H. Cherifi "Reduced reference 3D mesh quality assessment based on statistical models", in : Signal-Image Technology & Internet-Based Systems (SITIS), International Conference on. IEEE, November 2015 Bangkok-Thailande.
- [C2] **I. Abouelaziz**, M. EL Hassouni, "Evaluation de la qualité visuelle des objets 3D", Journée URAC, Rabat, 28 Novembre 2015.
- [C1] **I. Abouelaziz**, M. EL Hassouni, "New Models of Visual Saliency : Contourlet Transform Based Model and Hybrid Model", in IEEE Intelligent Systems and Computer Vision (ISCV), March 2015 Fez-Morocco.

Bibliographie

Abouelaziz, I., Chetouani, A., El Hassouni, M. et Cherifi, H. (2018). A blind mesh visual quality assessment method based on convolutional neural network. *Electronic Imaging*, 2018(18):423–1.

Abouelaziz, I., El Hassouni, M. et Cherifi, H. (2016a). A curvature based method for blind mesh visual quality assessment using a general regression neural network. *In Signal-Image Technology & Internet-Based Systems (SITIS), 2016 12th International Conference on*, pages 793–797. IEEE.

Abouelaziz, I., El Hassouni, M. et Cherifi, H. (2016b). No-reference 3d mesh quality assessment based on dihedral angles model and support vector regression. *In International Conference on Image and Signal Processing*, pages 369–377. Springer.

Abouelaziz, I., El Hassouni, M. et Cherifi, H. (2017). A convolutional neural network framework for blind mesh visual quality assessment. *In 2017 IEEE International Conference on Image Processing (ICIP)*, pages 755–759. IEEE.

Algashaam, F. M., Nguyen, K., Alkanhal, M., Chandran, V., Boles, W. et Banks, J. (2017). Multispectral periocular classification with multimodal compact multi-linear pooling. *IEEE Access*, 5:14572–14578.

Alliez, P. et Gotsman, C. (2005). Recent advances in compression of 3d meshes. *In Advances in multiresolution for geometric modelling*, pages 3–26. Springer.

Aspert, N., Santa-Cruz, D. et Ebrahimi, T. (2002). Mesh : Measuring errors between surfaces using the hausdorff distance. *In Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 705–708. IEEE.

Babu, R. V., Suresh, S. et Perkis, A. (2007). No-reference jpeg-image quality assessment using gap-rbf. *Signal Processing*, 87(6):1493–1503.

Bai, S., Bai, X., Zhou, Z., Zhang, Z., Tian, Q. et Latecki, L. J. (2017). Gift : Towards scalable 3d shape retrieval. *IEEE Transactions on Multimedia*, 19(6):1257–1271.

Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, 1(3):295–311.

Belkin, M., Niyogi, P. et Sindhwani, V. (2006). Manifold regularization : A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434.

Bian, Z., Hu, S.-M. et Martin, R. R. (2009). Evaluation for small visual difference between conforming meshes on strain field. *Journal of Computer Science and Technology*, 24(1):65–75.

Botsch, M., Kobbelt, L., Pauly, M., Alliez, P. et Lévy, B. (2010). *Polygon mesh processing*. CRC press.

Breitmeyer, B. G. (2007). Visual masking : past accomplishments, present status, future developments. *Advances in cognitive psychology*, 3(1-2):9.

Buades, A., Coll, B. et Morel, J.-M. (2005). A non-local algorithm for image denoising. *In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE.

Bui, T. N. et Jones, C. (1992). Finding good approximate vertex and edge partitions is np-hard. *Information Processing Letters*, 42(3):153–159.

Bulbul, A., Capin, T., Lavoué, G. et Preda, M. (2011). Assessing visual quality of 3-d polygonal models. *IEEE Signal Processing Magazine*, 28(6):80–90.

Bullier, J. (1998). Architecture fonctionnelle du système visuel. *Vision : aspects perceptifs et cognitifs*, pages 11–42.

Caruana, R. et Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *In Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

Chartier, S., Boukadoum, M. et Amiri, M. (2009). Bam learning of nonlinearly separable tasks by using an asymmetrical output function and reinforcement learning. *IEEE transactions on neural networks*, 20(8):1281–1292.

Chetouani, A. (2018a). Three-dimensional mesh quality metric with reference based on a support vector regression model. *Journal of Electronic Imaging*, 27(4):043048.

Chetouani, A. (2018b). Three-dimensional mesh quality metric with reference based on a support vector regression model. *Journal of Electronic Imaging*, 27(4):043048.

Chetouani, A., Beghdadi, A., Chen, S. et Mostafaoui, G. (2010). A novel free reference image quality metric using neural network approach. *In Proceedings of International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pages 1–4.

Chung, F. R. et Graham, F. C. (1997). *Spectral graph theory*. Numéro 92. American Mathematical Soc.

Cignoni, P., Rocchini, C. et Scopigno, R. (1998). Metro : Measuring error on simplified surfaces. *In Computer Graphics Forum*, volume 17, pages 167–174. Wiley Online Library.

Cohen-Steiner, D. et Morvan, J.-M. (2003). Restricted delaunay triangulations and normal cycle. *In Proceedings of the nineteenth annual symposium on Computational geometry*, pages 312–321.

Corsini, M., Gelasca, E. D., Ebrahimi, T. et Barni, M. (2007). Watermarked 3-d mesh quality assessment. *IEEE Transactions on Multimedia*, 9(2):247–256.

Daly, S. (2001). Engineering observations from spatiovelocity and spatiotemporal visual models. *In Vision Models and Applications to Image and Video Processing*, pages 179–200. Springer.

Defferrard, M., Bresson, X. et Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *In Advances in neural information processing systems*, pages 3844–3852.

Denis, G. (2014). *Apport de la vision par ordinateur dans l'utilisabilité des neuroprothèses visuelles.* Thèse de doctorat, Université de Toulouse, Université Toulouse III-Paul Sabatier.

Dhillon, I. S., Guan, Y. et Kulis, B. (2007). Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11):1944–1957.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. et Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. *In Advances in neural information processing systems*, pages 2224–2232.

Ebrahimi, T. (2009). Quality of multimedia experience : past, present and future. *In Proceedings of the 17th ACM international conference on Multimedia*, pages 3–4.

Elad, M. (2005). Retinex by two bilateral filters. *In International Conference on Scale-Space Theories in Computer Vision*, pages 217–229. Springer.

Engeldrum, P. G. (2000). *Psychometric scaling : a toolkit for imaging systems development.* Imcotek.

Engelke, U., Pépion, R., Le Callet, P. et Zepernick, H.-J. (2010). Linking distortion perception and visual saliency in h. 264/avc coded video containing packet loss. *In Visual Communications and Image Processing 2010*, volume 7744, page 774406. International Society for Optics and Photonics.

Feng, X., Wan, W., Da Xu, R. Y., Perry, S., Li, P. et Zhu, S. (2018). A novel spatial pooling method for 3d mesh quality assessment based on percentile weighting strategy. *Computers & Graphics*, 74:12–22.

FUKUI, A., PARK, D. H., YANG, D., ROHRBACH, A., DARRELL, T. et ROHRBACH, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv :1606.01847.*

GARLAND, M. et HECKBERT, P. S. (1997). Surface simplification using quadric error metrics. *In Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216. ACM Press/Addison-Wesley Publishing Co.

GASTALDO, P. et ZUNINO, R. (2005). Neural networks for the no-reference assessment of perceived quality. *Journal of Electronic Imaging*, 14(3):033004–033004.

GASTALDO, P., ZUNINO, R., HEYNDERICKX, I. et VICARIO, E. (2005). Objective quality assessment of displayed images by using neural networks. *Signal Processing : Image Communication*, 20(7):643–661.

GELASCA, E. D., EBRAHIMI, T., CORSINI, M. et BARNI, M. (2005). Objective evaluation of the perceptual quality of 3d watermarking. *In Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 1, pages I–241. IEEE.

GIUSTI, A., CIREŞAN, D. C., MASCI, J., GAMBARDELLA, L. M. et SCHMIDHUBER, J. (2013). Fast image scanning with deep max-pooling convolutional neural networks. *In 2013 IEEE International Conference on Image Processing*, pages 4034–4038. IEEE.

GORI, M., MONFARDINI, G. et SCARSELLI, F. (2005). A new model for learning in graph domains. *In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE.

GROVER, A. et LESKOVEC, J. (2016). node2vec : Scalable feature learning for networks. *In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

HAMMOND, D. K., VANDERGHEYNST, P. et GRIBONVAL, R. (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150.

HASSOUN, M. H. *et al.* (1995). *Fundamentals of artificial neural networks.* MIT press.

HE, K., ZHANG, X., REN, S. et SUN, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

HOWARD, I. P. et ROGERS, B. J. (2002). *Seeing in depth, Vol. 2 : Depth perception.* University of Toronto Press.

HSU, C.-W., CHANG, C.-C., LIN, C.-J. *et al.* (2003). A practical guide to support vector classification.

HUBEL, D. H. et WIESEL, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154.

Johnson, A. E. et Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on pattern analysis and machine intelligence*, 21(5):433–449.

Karni, Z. et Gotsman, C. (2000). Spectral compression of mesh geometry. *In Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 279–286. ACM Press/Addison-Wesley Publishing Co.

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv :1404.5997*.

Krizhevsky, A., Sutskever, I. et Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems*, pages 1097–1105.

Lavoué, G. (2009). A local roughness measure for 3d meshes and its application to visual masking. *ACM Transactions on Applied perception (TAP)*, 5(4):21.

Lavoué, G. (2011). A multiscale metric for 3d mesh visual quality assessment. *In Computer Graphics Forum*, volume 30, pages 1427–1437. Wiley Online Library.

Lavoué, G. et Corsini, M. (2010). A comparison of perceptually-based metrics for objective evaluation of geometry processing. *IEEE Transactions on Multimedia*, 12(7): 636–649.

Lavoué, G., Gelasca, E. D., Dupont, F., Baskurt, A. et Ebrahimi, T. (2006). Perceptually driven 3d distance metrics with application to watermarking. *In SPIE Optics+ Photonics*, pages 63120L–63120L. International Society for Optics and Photonics.

LeCun, Y., Bengio, Y. et Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

LeCun, Y., Bottou, L., Bengio, Y. et Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, C. H., Varshney, A. et Jacobs, D. W. (2005). Mesh saliency. *In ACM SIGGRAPH 2005 Papers*, pages 659–666.

Lee, H., Dikici, Ç., Lavoué, G. et Dupont, F. (2011). Joint reversible watermarking and progressive compression of 3d meshes. *The Visual Computer*, 27(6):781–792.

Leifman, G., Shtrom, E. et Tal, A. (2012). Surface regions of interest for viewpoint selection. *In 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 414–421. IEEE.

Li, C., Bovik, A. C. et Wu, X. (2011a). Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks*, 22(5):793–799.

Li, C., Bovik, A. C. et Wu, X. (2011b). Blind image quality assessment using a general regression neural network. *IEEE Transactions on neural networks*, 22(5):793–799.

Li, Y., Tarlow, D., Brockschmidt, M. et Zemel, R. (2015). Gated graph sequence neural networks. *arXiv preprint arXiv :1511.05493*.

Lin, W., Ebrahimi, T., Loizou, P. C., Moller, S. et Reibman, A. R. (2012). Introduction to the special issue on new subjective and objective methodologies for audio and visual signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):614–615.

Lin, W. et Kuo, C.-C. J. (2011). Perceptual visual quality metrics : A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312.

Luebke, D. P. (2001). A developer's survey of polygonal simplification algorithms. *IEEE Computer Graphics and Applications*, 21(3):24–35.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems*, pages 3111–3119.

Mittal, A., Moorthy, A. K. et Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12): 4695–4708.

Moorthy, A. K. et Bovik, A. C. (2010). A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters*, 17(5):513–516.

Nair, V. et Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *In Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Narwaria, M. et Lin, W. (2010). Objective image quality assessment based on support vector regression. *IEEE Transactions on Neural Networks*, 21(3):515–519.

Nouri, A., Charrier, C. et Lézoray, O. (2017). 3d blind mesh quality assessment index. *Electronic Imaging*, 2017(20):9–26.

Pan, Y., Cheng, I. et Basu, A. (2005). Quality metric for approximating subjective evaluation of 3-d objects. *IEEE Transactions on Multimedia*, 7(2):269–279.

Papandreou, G., Kokkinos, I. et Savalle, P.-A. (2015). Modeling local and global deformations in deep learning : Epitomic convolution, multiple instance learning, and sliding window detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 390–399.

Parkhurst, D., Law, K. et Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123.

Perozzi, B., Al-Rfou, R. et Skiena, S. (2014). Deepwalk : Online learning of social representations. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

Rogowitz, B. E. et Rushmeier, H. E. (2001). Are image quality metrics adequate to evaluate the quality of geometric objects ? *In Human Vision and Electronic Imaging VI*, volume 4299, pages 340–348. International Society for Optics and Photonics.

Rushmeier, H. E., Rogowitz, B. E. et Piatko, C. (2000). Perceptual issues in substituting texture for geometry. *In Human Vision and Electronic Imaging V*, volume 3959, pages 372–383. International Society for Optics and Photonics.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. *et al.* (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Saad, M. A., Bovik, A. C. et Charrier, C. (2010). A dct statistics-based blind image quality index. *IEEE Signal Processing Letters*, 17(6):583–586.

Sarle, W. S. (1996). Stopped training and other remedies for overfitting. *Computing science and statistics*, pages 352–360.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. et Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Silva, S., Santos, B. S., Ferreira, C. et Madeira, J. (2009). A perceptual data repository for polygonal meshes. *In 2009 Second International Conference in Visualisation*, pages 207–212. IEEE.

Silva, S., Santos, B. S., Madeira, J. et Ferreira, C. (2008). Perceived quality assessment of polygonal meshes using observer studies : A new extended protocol. *In Human Vision and Electronic Imaging XIII*, volume 6806, page 68060D. International Society for Optics and Photonics.

Simonyan, K. et Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556.*

Song, R., Liu, Y., Martin, R. R. et Rosin, P. L. (2014). Mesh saliency via spectral processing. *ACM Transactions on Graphics (TOG)*, 33(1):1–17.

Song, R., Liu, Y., Zhao, Y., Martin, R. R. et Rosin, P. L. (2012). Conditional random field-based mesh saliency. *In 2012 19th IEEE International Conference on Image Processing*, pages 637–640. IEEE.

Sorkine, O., Cohen-Or, D. et Toledo, S. (2003). High-pass quantization for mesh encoding. *In Symposium on Geometry Processing*, volume 42.

Specht, D. F. *et al.* (1991). A general regression neural network. *IEEE transactions on neural networks*, 2(6):568–576.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. et Mei, Q. (2015). Line : Large-scale information network embedding. *In Proceedings of the 24th international conference on world wide web*, pages 1067–1077.

Tatler, B. W., Baddeley, R. J. et Gilchrist, I. D. (2005). Visual correlates of fixation selection : Effects of scale and time. *Vision research*, 45(5):643–659.

Torkhani, F., Wang, K. et Chassery, J.-M. (2012). A curvature tensor distance for mesh visual quality assessment. *Computer Vision and Graphics*, pages 253–263.

Torkhani, F., Wang, K. et Chassery, J.-M. (2014). A curvature-tensor-based perceptual quality metric for 3d triangular meshes. *Machine Graphics & Vision*, 23(1).

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Váša, L. et Rus, J. (2012). Dihedral angle mesh error : a fast perception correlated distortion measure for fixed connectivity triangle meshes. *In Computer Graphics Forum*, volume 31, pages 1715–1724. Wiley Online Library.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wang, K., Lavoué, G., Denis, F. et Baskurt, A. (2008). A comprehensive survey on three-dimensional mesh watermarking. *IEEE Transactions on Multimedia*, 10(8):1513–1527.

Wang, K., Torkhani, F. et Montanvert, A. (2012). A fast roughness-based approach to the assessment of 3d mesh visual quality. *Computers & Graphics*, 36(7):808–818.

Wang, Y.-P. et Hu, S.-M. (2009). A new watermarking method for 3d models based on integral invariants. *IEEE Transactions on Visualization and Computer Graphics*, 15(2):285–294.

Wang, Z. et Bovik, A. C. (2006). Modern image quality assessment (synthesis lectures on image, video, and multimedia processing). *San Rafael, CA : Morgan Claypool.*

Wang, Z. et Bovik, A. C. (2009). Mean squared error : Love it or leave it ? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117.

Wang, Z., Bovik, A. C., Sheikh, H. R. et Simoncelli, E. P. (2004). Image quality assessment : from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Watson, B., Friedman, A. et McGaffey, A. (2001). Measuring and predicting visual fidelity. *In Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 213–220.

Weston, J., Ratle, F., Mobahi, H. et Collobert, R. (2012). Deep learning via semi-supervised embedding. *In Neural networks : Tricks of the trade*, pages 639–655. Springer.

Wolfe, J. M., Alvarez, G. A. et Horowitz, T. S. (2000). Attention is fast but volition is slow. *Nature*, 406(6797):691–691.

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5:23.

Wu, J., Shen, X., Zhu, W. et Liu, L. (2013). Mesh saliency with global rarity. *Graphical Models*, 75(5):255–264.

Wu, J.-H., Hu, S.-M., Tai, C.-L. et Sun, J.-G. (2001). An effective feature-preserving mesh simplification scheme based on face constriction. *In Proceedings Ninth Pacific Conference on Computer Graphics and Applications. Pacific Graphics 2001*, pages 12–21. IEEE.

Yang, Z., Cohen, W. W. et Salakhutdinov, R. (2016). Revisiting semi-supervised learning with graph embeddings. *arXiv preprint arXiv :1603.08861.*

Yarbus, A. L. (2013). *Eye movements and vision.* Springer.

Yildiz, Z. C. et Capin, T. (2017). A perceptual quality metric for dynamic triangle meshes. *EURASIP Journal on Image and Video Processing*, 2017(1):12.

Zeiler, M. D. et Fergus, R. (2014). Visualizing and understanding convolutional networks. *In European conference on computer vision*, pages 818–833. Springer.

Zhang, W., Qu, C., Ma, L., Guan, J. et Huang, R. (2016). Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. *Pattern Recognition*, 59:176–187.

Zhao, Y., Liu, Y., Song, R. et Zhang, M. (2012). Extended non-local means filter for surface saliency detection. *In 2012 19th IEEE International Conference on Image Processing*, pages 633–636. IEEE.

Zhao, Y., Liu, Y. et Zeng, Z. (2013). Using region-based saliency for 3d interest points detection. *In International Conference on Computer Analysis of Images and Patterns*, pages 108–116. Springer.

Zhu, X., Ghahramani, Z. et Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *In Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919.

Zhu, Z., Rao, C., Bai, S. et Latecki, L. J. (2017). Training convolutional neural network from multi-domain contour images for 3d shape retrieval. *Pattern Recognition Letters.*

ROYAUME DU MAROC
Université Mohammed V جامعة محمد الخامس
- RABAT - – الرباط –
Faculté des sciences
كلية العلوم

CENTRE D'ETUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

Résumé

De même que les images et les vidéos, la qualité perceptuelle des maillages 3D peut être affectée par plusieurs facteurs externes (tatouage, compression, simplification, etc.). Afin d'estimer l'impact de ces traitements, plusieurs mesures de qualité ont été proposées dans la littérature. Ils existent actuellement trois grandes familles de métriques : référence complète (FR) qui suppose que le maillage de référence est disponible, référence réduite (RR) qui exploite uniquement des caractéristiques du maillage de référence et les mesures sans référence (NR), appelées également aveugle, qui n'ont accès à aucune information du maillage de référence. Un certain nombre de méthodes à référence complètes ou réduites ont été proposées afin d'estimer la qualité visuelle perçue des maillages 3D. Cependant, dans la plupart des situations pratiques, l'accès aux informations relatives à la référence et au type de distorsion est limité. Pour ces raisons, le développement d'une méthode de qualité visuelle sans référence est une problématique cruciale. Dans ce cadre, nous nous intéressons dans cette thèse à l'évaluation sans référence de la qualité perceptuelle des maillages 3D. La première contribution porte sur la proposition d'une méthode basée sur une approche d'apprentissage afin de prédire les scores de qualité. Cette méthode utilise des caractéristiques extraites à partir du maillage déformé et des méthodes d'apprentissage pour l'estimation de la qualité visuelle. Quant à la seconde contribution, l'apprentissage profond est exploité pour estimer la qualité. Nous commençons par un réseau CNN simple et ensuite nous utilisons des réseaux plus profonds avec une combinaison basée sur le Compact Multi-linear Pooling (CMP). Dans la troisième contribution la saillance visuelle 3D est exploitée pour préparer les données d'apprentissage. Les architectures CNN sont alimentées par des petits patchs soigneusement sélectionnés en fonction de leur niveau de saillance. Dans la dernière contribution, nous introduisons une méthode "model-based» utilisant des réseaux convolutionnels de graphes (GCN) pour traiter directement le modèle 3D lui-même sans utiliser des patchs rendus (image-based).

Mots-clefs (8) : Maillage triangulaire. Évaluation sans référence de la qualité des maillage 3D. Système visuel humain. Évaluation objective. Évaluation subjective. Qualité perceptuelle. L'apprentissage profond. Saillance visuelle. Réseau de neurones convolutifs.

Abstract

As like images and videos, the perceptual quality of 3D meshes can be affected by several external factors (watermarking, compression, simplification, etc.). To estimate the impact of these treatments, several quality measures have been proposed in the literature. There are currently three main families of metrics: full reference (FR) which assumes that the reference mesh is available, reduced reference (RR) which only uses extracted features from the reference mesh, and no reference measures (NR), also called blind, which do not have access to any information of reference. A certain full or reduced reference methods have been proposed to estimate the perceived visual quality of 3D meshes. However, in most practical situations, access to information about the reference and the type of distortion is limited. For these reasons, the development of a no-reference method is a crucial issue. In this context, we are interested in this thesis in no-reference mesh visual quality assessment. The first contribution concerns the development of a method based on feature learning approaches to predict the quality scores. This method uses features extracted from the distorted mesh and feature learning methods for quality estimation. As for the second contribution, deep learning is used to assess the visual quality. We begin by a simple CNN network from scratch and then we take advantage of deeper networks with a combination based on the Compact Multi-Linear Pooling (CMP). In the third contribution, the 3D visual saliency is used to prepare the learning data. CNN architectures are fed by small patches carefully selected according to their level of saliency. In the last contribution, we introduce a "model-based" method using convolutional graph networks (GCN) to directly process the 3D model itself without using 2D patches rendered from the 3D model (image-based).

.

Key Words (8): Triangular mesh. Blind mesh quality assessment. Human visual system. Objective evaluation. Subjective evaluation. Perceptual quality. Deep learning. Visual saliency. Convolutional neural network.