A CONVOLUTIONAL NEURAL NETWORK FRAMEWORK FOR BLIND MESH VISUAL QUALITY ASSESSMENT

Ilyass Abouelaziz[1], *Mohammed El Hassouni*[1] *and Hocine Cherifi*[2]

[1]LRIT, Associated Unit to CNRST (URAC No 29)- Faculty of Sciences,
Mohammed V University in Rabat, B.P.1014 RP, Rabat, Morocco.
[2]LE2I UMR 6306 CNRS, University of Burgundy, Dijon, France.

ABSTRACT

In this paper, we propose a new method for blind mesh visual quality assessment using a deep learning approach. To do this, we first extract visual representative features by computing locally curvature and dihedral angles from each distorted mesh. Then, we determine from these features a set of 2D patches which are learned to a convolutional neural network (CNN). The network consists of two convolutional layers with two max-pooling layers. Then, a multilayer perceptron (MLP) with two fully connected layers is integrated to summarize the learned representation into an output node. With this network structure, feature learning and regression are used to predict the quality score of a given distorted mesh without needing to a reference mesh. Experiments are conducted on LIRIS masking and the general-purpose databases and results show that the trained CNN achieves good rates in terms of correlation with human visual judgment scores.

Index Terms— Convolutional neural network (CNN), blind mesh visual quality assessment, Human visual system, mean curvature, dihedral angles.

1. INTRODUCTION

The low perceived visual quality of 3D meshes is a result of different lossy operations related to the transmission and geometric processing, such as watermarking, simplification and so forth [1, 2]. The perceptual quality of a 3D mesh is usually defined subjectively as the mean of the individual ratings by human subjects (MOS: Mean opinion score). However, subjective quality assessment is too expensive, laborious and time-consuming. Objective visual quality assessment methods are the ultimate solution to automatically assess the visual quality [3]. The problem of assessing the visual quality of 3D meshes has known a considerable progress in the last years. The early work used simple similarities between the reference mesh and its distorted version such as root mean square error (RMS) [4] and Hausdorff distance (HD) [5]. This kind of methods generally failed to reflect the perceived visual quality since it computes a pure geometric distance neglecting the main operations of the human visual system (HVS) [6]. In order to incorporate the perceptual information, several methods use different perceptual principles for a better estimation of the perceived quality. In [7], a perceptual metric based on the curvature analysis called mesh structural distortion measure (MSDM) has been proposed. In order to evaluate the quality of watermarked meshes, Corsini et *al.* developed a perceptual metric using the roughness variation [8]. Another perceptual metric called the fast mesh perceptual distance (FMPD) has been proposed in [9]. This metric is based on a mesh local roughness measure derived from Gaussian curvature. These methods are full reference and achieve a very high correlation with human perception. However, their main drawback is the non-availability of the reference mesh in real world applications.

The ability to automatically predict the perceived quality is a challenging issue, especially in many practical computer vision applications when the reference is not available. Given only the distorted mesh, the no-reference approach tries to predict the perceived visual quality and ensures a good correlation with human judgments without taking into consideration the reference models. This concept is extensively addressed in image quality assessment, and several methods successfully estimate the quality score of distorted images by taking advantage of extracted features and exploiting machine learning methods in order to learn the extracted features and provide the quality score [10]. However, in mesh quality assessment there is a distinct lack of no-reference methods, and exploiting this concept would be an important gain, especially with the promising results obtained in our previous work concerning the blind quality assessment using support vector regression (SVR) [11]. Motivated by the above observations, we propose in this paper a deep learning based method for blind mesh visual quality assessment. Recently, deep learning has been widely used for no-reference image quality assessment task [12, 13] and achieved great performance thanks to its ability to learn discriminant features.

The proposed method uses 2D patches computed from curvature and dihedral angles extracted features, and a CNN based learning approach to predict the perceived visual quality of distorted 3D meshes with different distortion types.

This paper is organized as follows. Section. 2 presents the learning framework with a convolutional neural network (CNN). In Section. 3, we describe the proposed method. Experimental results and discussion are given in Section. 4. Finally, conclusions are drawn in Section. 5.

2. LEARNING FRAMEWORK WITH CONVOLUTIONAL NEURAL NETWORK (CNN)

Deep learning has attracted the attention of many researchers and reached high performances on various computer vision applications [14]. Specifically, CNN has shown great success on image processing thanks to its suitability and degree of freedom. In this section, we present the different layers of a CNN.

2.1. Convolution

Convolution is defined as the process of filtering through the input patch to look for a specific pattern. The convolution performance is mainly affected by two types of parameters called weights and biases. We denote X the input patch of the CNN. $\{W_i\}_{i=1}^{N}$ are N convolutional kernels and $\{b_i\}_{i=1}^{N}$ are the biases values. The convolution process can be defined as follows:

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

Where $*$ indicates the convolution operation. After the convolution process, N features maps $\{Y_i\}_{i=1}^{N}$ are generated. We note that the parameters $\{W_i\}_{i=1}$ and $\{b_i\}_{i=1}$ are shared for each convolution layer.

2.2. Pooling

After the convolution process, multiple feature maps are obtained with a richer representation. The next step is to apply a pooling process on each feature map to reduce the filter responses to a lower dimension. In this work, we tend to use the max-pooling process. The feature map is partitioned into a set of rectangles according to a local window, and provide the maximum value for each sub-region. Let $Y_{x,y}^{n}$ denote the response at location (x, y) of the feature map obtained by the n^{th} filter. The max-pooling process is defined as:

$$M_{x,y}^{n} = \max_{(x,y)\in\Omega}(Y_{x,y}^{n}) \tag{2}$$

Where $n = 1, 2, ..., N$ and N is the number of kernels. $M_{x,y}^{n}$ represents the maximum values obtained after the pooling process. Ω denotes the local window for the pooling process. If the size of the local window is equal to the size of the feature map, the pooling method reduces the feature maps to a one-dimensional feature vector.

2.3. Multilayer perception (MLP)

After the convolution and the pooling process, a multilayer perception (MLP) with two fully connected layers is used for the final representation. In the quality assessment, MLP is used to summarize the representation obtained by the previous layers and generate the quality score as follows:

$$Q_s = \omega_s \sigma + b_s \tag{3}$$

where ω_s and b_s are the learned parameters to compute the quality score of the input patch. Q_s is the predicted quality score that indicates the perceived perceptual quality of a given input patch. σ is the non-linear activation function. In this work, we tend to use the Rectified Linear Units (ReLU) as a non-linear activation function in the two fully connected layers. The ReLU activation function is defined as:

$$\sigma = \max\left(0, \sum_i \omega_i a_i\right) \tag{4}$$

Where ω_i denotes the weights of the ReLU and a_i is the output of the previous layer (pooling layer). The ReLU is used in the two fully connected layers instead of the traditional sigmoid and $\tanh$ neurons.

3. PROPOSED MESH VISUAL QUALITY ASSESSMENT METHOD WITH CNN

3.1. Methodology

The different steps of the proposed mesh visual quality assessment method are depicted in Fig. 1. First, two perceptual features i.e mean curvature and dihedral angles are extracted. Then, the extracted features are reorganized into 2-D small patches in order to fit the input of the CNN. After that, a CNN architecture is proposed followed by a regression method and the trained regression model estimates the quality score for each patch. Finally, the overall quality score is obtained by averaging all predicted patch scores.

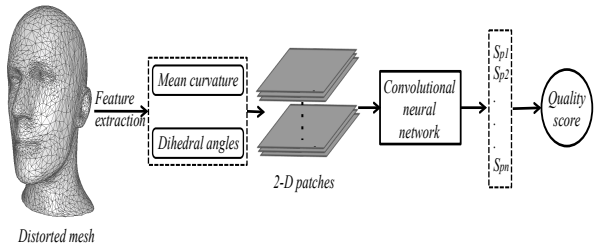

Fig. 1. Overall scheme of the proposed method.

3.2. Feature extraction and preparation of 2-D patches

To learn a convolutional neural network, several image processing studies work in the spatial domain by splitting the

image into small patches. Afterward, these patches are used as inputs for the network. In our work, we aim to learn a compact and effective mesh representation from low-level features. Therefore, two types of perceptual features are extracted: mean curvature and dihedral angles. The mean curvature is an important perceptual feature representing the visual aspect of a 3D mesh. It describes the deviation amount of a surface of being flat and provides several visual characteristics of a 3D model, particularly sharpness, roughness or smoothness of a region. The structural aspect of the 3D mesh is represented by the dihedral angle which is used to construct the concept of global roughness. These geometric features are widely used in many mesh processing applications [15, 9] and can describe the 3D mesh from multiple perspectives. The extracted low-level features can be concatenated into a high dimensional feature vector. However, this representation may lead to the over-fitting due to the high-dimensional descriptor space. In addition, it does not fit the convolutional property of CNNs. Thus, we propose to reorganize the high-dimensional feature vector into 2-D small patches. This representation allows using the convolutional property of the CNN.

3.3. CNN configuration

The architecture of our proposed CNN is composed by seven layers as depicted in Fig. 2.

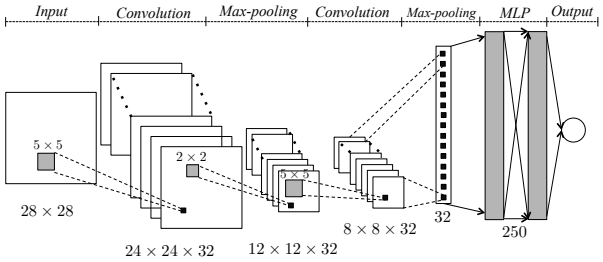

Fig. 2. Convolutional neural network configuration for mesh quality assessment.

The different layers are presented as follows:

- **Input:** 2-D patches of size 28×28.
- **Convolutional layer 1:** The first layer is a convolutional layer which filters the input patch with 32 kernels. Each kernel is of size (5×5). This layer provides 32 feature maps of size 24×24.
- **Max-pooling layer 1:** The second layer is a max-pooling layer which applies the max-pooling process on each feature map with a local window of size 2×2. This layer produces 32 feature maps with a lower dimension of 12×12.
- **Convolutional layer 2:** The third layer is another convolutional layer which filters the output of the max-pooling layer with 32 kernels of size 5×5. This layer produces 32 feature maps of size 8×8.
- **Max-pooling layer 2:** The fourth layer is another max-pooling layer with a local window of size 8×8. as a result, this layer produces a feature vector of size 1×32.
- **Fully connected layers:** The fifth and sixth layers are two fully connected layers of 250 nodes each.
- **Output layer:** The seventh and last layer is a simple linear regression with a one-dimensional output that provides the quality score.

3.4. Training and quality prediction

Our network is trained on non-overlapping 2-D patches of size (28×28) obtained from the extracted features of the 3D meshes. Thanks to the homogeneous distortions in the training meshes, we assign for each input patch a score the same as the mean opinion score of the source mesh. Similarly to [16], we adopt the training objective function defined as follows:

$$L = \frac{1}{N} \sum_{n=1}^{N} ||S(p_n; \omega) - MOS_n||_{l1} \qquad (5)$$

$$\hat{\omega} = \min_{\omega} L$$

Where MOS_n is the mean opinion score assigned to a given input patch p_n and $S(p_n; \omega)$ is the predicted score of p_n with network weights ω. The parameters of the convolutional neural network are learned using stochastic gradient descent (SGD) and back propagation by minimizing the objective function defined in Eq. 5. In our experiments, we perform stochastic gradient descent for 40 epochs.

In the test phase, the quality score for a given patch is obtained using the model parameters that ensure the best correlation with the mean opinion scores. Finally, The overall quality score for a specific distorted mesh is the average of all predicted patch scores.

4. EXPERIMENTAL RESULTS

4.1. Datasets and validation protocol

To test the performance of a mesh visual quality MVQ method, a dataset of distorted meshes graded by human observers is needed. Our blind MVQ assessment method has been tested and validated using two publicly available datasets specially designed for quality methods evaluation.

- LIRIS Masking database [17] that contains 4 reference models and 24 distorted models obtained by the local noise addition with different levels.
- General-purpose database [18] that contains 4 reference models and 84 distorted models obtained by the local noise addition and smoothing with different levels.

Table 1. Correlation coefficients r_s (%) and r_p (%) of different objective methods.

Database	Methods	HD [5]		RMS [4]		3DWPM2 [8]		MSDM2 [7]		FMPD [9]		Proposed method	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
LIRIS masking	Armadillo	48.6	37.7	65.6	44.6	48.6	37.9	81.1	88.6	94.2	88.6	95.2	97.6
	Lion	71.4	25.1	71.4	23.8	38.3	22.0	93.5	94.3	93.5	94.3	89.4	91.6
	Bimba	25.7	7.5	71.4	21.8	37.1	14.4	96.8	100	98.9	100	93.4	98.7
	Dyno	48.6	31.1	71.4	50.3	71.4	50.1	95.6	100	96.9	94.3	96.3	89.9
	Whole repository	26.6	4.1	48.8	17.0	37.4	18.2	87.3	89.6	80.8	80.2	88.2	85.4
General purpose	Armadillo	69.5	30.2	62.7	32.3	74.1	43.1	81.6	85.3	75.4	83.2	87.2	84.3
	Dyno	30.9	22.6	0.3	0.0	52.4	19.9	85.4	85.7	89.6	88.9	86.4	86.2
	Venus	1.6	0.8	90.1	77.3	34.8	16.4	89.3	87.5	87.5	83.9	92.2	85.6
	Rocker	18.1	5.5	7.3	3.0	37.8	29.9	89.6	87.2	88.8	84.7	91.3	85.2
	Whole repository	13.8	1.3	26.8	7.9	49.0	24.6	80.4	81.4	81.9	83.5	83.6	82.7

We note that the mesh representation used in our method provides an important number of 2D patches that make the dataset huge enough for the training process.

To evaluate the performance of the proposed method, two correlation coefficients are commonly used, namely, the Pearson linear correlation coefficient (r_p) to measure the prediction accuracy, and the Spearman rank-order correlation coefficient (r_s) to measure the prediction monotonicity [19]. Since the obtained scores and the mean opinion scores are non-linear, it is highly recommended to introduce a psychometric fitting in order to partially remove this non-linearity. We note that in this work, we adopt a cumulative Gaussian psychometric function as defined in [20].

4.2. Results and discussion

To evaluate the performance of the proposed blind mesh quality assessment method, a comparison has been done with methods reported in the literature that are HD[5], RMS[4], 3DWPM2[8], MSDM2[7] and FMPD[9]. Table.1 reports the obtained correlation coefficients r_s and r_p of the compared methods on the two considered databases. The correlations on the whole corpus are computed between the objective scores of all objects in the corpus and their corresponding MOSs. We can remark that the methods based on geometric distance HD and RMS generally fail to reflect the perceived visual quality and do not correlate well with human perception. The principle of these methods is to compute a pure geometric distance. Thus, the main operations of the human visual system are neglected. On the other hand, perceptually driven methods MSDM2, FMPD, 3DWPM2 and the proposed method achieve high correlations with human judgments and show a good performance in assessing the perceived quality.

Concerning the LIRIS masking database, we can notice that the proposed method has the highest r_s score (88.2%) and the second highest r_p score (85.4%) regarding the whole repository. Thus, it outperforms two of the most influential and effective methods in the-state-of-the-art that are MSDM2 and FMPD. In addition, our method produces high correlations for each mesh individually, notably the Armadillo mesh with the highest r_s (95.2%) and r_p (97.6%) among the other methods.

The good performance of the proposed method can be also proved by the competitive scores on the general-purpose database. Particularly, the highest r_s on the whole repository (83.6%) and the highest scores for Armadillo, Venus and Rocker models. Comparing to the LIRIS masking database, the general purpose database contains an important number of distorted models (21 distorted version for each model as well as a variety of distortion types). The high correlation scores provided by the proposed method in this database appear to be a good indicator for the forcefulness of our method in mesh visual quality assessment.

5. CONCLUSION

In this paper, we presented a convolutional neural network (CNN) for blind mesh visual quality assessment. The network is fed by perceptual features extracted from the 3D meshes and arranged into 2-D small patches to meet the requirements of the CNN. The proposed architecture is composed of multiple layers of convolution and max-pooling. In addition, the MLP layer with two fully connected layers is integrated to summarize the representation and produce the final quality score. The experimental results proved that the trained network successfully predicts the visual quality. In addition, it is noteworthy that the proposed method is blind and does not require the reference mesh. Unlike the competing full and reduced reference methods, our method can be useful in practical situations. Future work will concern other representative features and network configuration.

Acknowledgment. The research leading to these results has received funding from the Regional Council of Burgundy, France.

6. REFERENCES

[1] Kai Wang, Guillaume Lavoué, Florence Denis, and Atilla Baskurt, "A comprehensive survey on three-dimensional mesh watermarking," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1513–1527, 2008.

[2] Michael Garland and Paul S. Heckbert, "Surface simplification using quadric error metrics," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1997, SIGGRAPH '97, pp. 209–216, ACM Press/Addison-Wesley Publishing Co.

[3] G. Lavoué and M. Corsini, "A comparison of perceptually-based metrics for objective evaluation of geometry processing," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 636–649, Nov 2010.

[4] Paolo Cignoni, Claudio Rocchini, and Roberto Scopigno, "Metro: Measuring error on simplified surfaces," Tech. Rep., Paris, France, France, 1996.

[5] N. Aspert, D. Santa-Cruz, and T. Ebrahimi, "Mesh: measuring errors between surfaces using the hausdorff distance," in *Proceedings. IEEE International Conference on Multimedia and Expo*, 2002, vol. 1, pp. 705–708 vol.1.

[6] Bruno G Breitmeyer, "Visual masking: past accomplishments, present status, future developments," *Advances in cognitive psychology*, vol. 3, no. 1-2, pp. 9–20, 2007.

[7] Guillaume Lavoué, Elisa Drelie Gelasca, Florent Dupont, Atilla Baskurt, and Touradj Ebrahimi, "Perceptually driven 3d distance metrics with application to watermarking," in *SPIE Optics+ Photonics*. International Society for Optics and Photonics, 2006, pp. 63120L–63120L.

[8] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, "Watermarked 3-d mesh quality assessment," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 247–256, Feb 2007.

[9] Kai Wang, Fakhri Torkhani, and Annick Montanvert, "A fast roughness-based approach to the assessment of 3d mesh visual quality," *Computers & Graphics*, vol. 36, no. 7, pp. 808–818, 2012.

[10] C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Transactions on Neural Networks*, vol. 22, no. 5, pp. 793–799, May 2011.

[11] Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi, *No-Reference 3D Mesh Quality Assessment Based on Dihedral Angles Model and Support Vector Regression*, pp. 369–377, Springer International Publishing, Cham, 2016.

[12] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1275–1286, June 2015.

[13] Wei Zhang, Chenfei Qu, Lin Ma, Jingwei Guan, and Rui Huang, "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network," *Pattern Recognition*, vol. 59, pp. 176 – 187, 2016, Compositional Models and Structured Learning for Visual Recognition.

[14] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 5 2015.

[15] Ran Gal and Daniel Cohen-Or, "Salient geometric features for partial shape matching and similarity," *ACM Trans. Graph.*, vol. 25, no. 1, pp. 130–150, Jan. 2006.

[16] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

[17] Guillaume Lavoué, Elisa Drelie Gelasca, Florent Dupont, Atilla Baskurt, and Touradj Ebrahimi, "Perceptually driven 3d distance metrics with application to watermarking," in *SPIE Optics+ Photonics*. International Society for Optics and Photonics, 2006, pp. 63120L–63120L.

[18] Guillaume Lavoué, "A local roughness measure for 3d meshes and its application to visual masking," *ACM Trans. Appl. Percept.*, vol. 5, no. 4, pp. 21:1–21:23, Feb. 2009.

[19] Zhou Wang and Alan C. Bovik, *Modern Image Quality Assessment*, Morgan and Claypool, 2006.

[20] Peter G Engeldrum, *Psychometric scaling: a toolkit for imaging systems development*, Imcotek, 2000.