CLUSTERING-BASED PSYCHOMETRIC NO-REFERENCE QUALITY MODEL FOR POINT CLOUD VIDEO

Sam Van Damme, Maria Torres Vega, Jeroen van der Hooft and Filip De Turck

IDLab, Department of Information Technology (INTEC)
Ghent University - imec
{firstname}.{lastname}@ugent.be

ABSTRACT

Point cloud video streaming is a fundamental application of immersive multimedia. In it, objects represented as sets of points are streamed and displayed to remote users. Given the high bandwidth requirements of this content, small changes in the network and/or encoding can affect the users' perceived quality in unexpected manners. To tackle the degradation of the service as fast as possible, real-time Quality of Experience (QoE) assessment is needed. As subjective evaluations are not feasible in real time due to their inherent costs and duration, low-complexity objective quality assessment is a must. Traditional No-Reference (NR) objective metrics at client side are best suited to fulfill the task. However, they lack on accuracy to human perception. In this paper, we present a cluster-based objective NR QoE assessment model for point cloud video. By means of Machine Learning (ML)-based clustering and prediction techniques combined with NR pixel-based features (e.g., blur and noise), the model shows high correlations (up to a 0.977 Pearson Linear Correlation Coefficient (PLCC)) and low Root Mean Squared Error (RMSE) (down to 0.077 on a zero-to-one scale) towards objective benchmarks after evaluation on an adaptive streaming point cloud dataset consisting of sixteen source videos and 453 sequences in total.

***Index Terms*—** Point clouds, Quality of Experience, objective quality metrics, psychometric curve-fitting, quality modelling

1. INTRODUCTION

Point cloud video is one of the promising applications in immersive multimedia. In point cloud delivery, objects composed by a dense network of 6D points (referring to geometry (e.g., x, y, z) and texture (e.g., the three color channels)) are presented to the remote user's Head-Mounted Display (HMD) or, alternatively, projected on a 2D screen. Stringent requirements in terms of bandwidth, latency and encoding can result in low quality rendering, therefore reducing the user's perceived quality (i.e., Quality of Experience (QoE)).

As subjective evaluations are costly in terms of time and money, they are not suited for real-time evaluation. Thus, objective quality assessments are needed. To objectively evaluate the quality of point clouds, two types of approaches exist: (i) geometric (quality of the point cloud object itself) and (ii) projection-based metrics (quality of the rendered video) [1]. While geometric metrics give an indication of the quality of the point cloud as a whole, they fail to capture the quality of the video actually rendered and displayed to the user. For this reason, traditional video quality metrics have recently been investigated as projection-based metrics, assessing the quality of the projected Field of View (FoV) [2]. They can be divided in Full-Reference (FR) (full comparison between the original and distorted sequence), No-Reference (NR) (quality assessment purely on the distorted received stream) and Reduced-Reference (RR) metrics (where a number of low-complexity features are sent over a side channel to the client for comparison [3]).

Within this type of research, multiple works are worth mentioning, of which the most prominent are discussed below. Yang *et al.* presented a FR metric based on the projection of the point cloud on the six perpendicular planes of a cube [4]. Combining features values (e.g., color, depth, texture and edges) from all six planes, a single quality index is derived. Results show Pearson Linear Correlation Coefficients (PLCCs) towards subjective Mean Opinion Scores (MOSs) ranging from 0.66 to 0.97, depending on the considered content and the introduced encoding distortions. Diniz *et al.* derived a RR point cloud quality assessment model based on local patterns [5]. In this model, each pixel is assigned a binary code by thresholding the difference in intensity with its surrounding pixels. The quality of the point cloud is then determined by the difference between the histograms of the original and the distorted content, mapping this distance to a predicted MOS using a third-order polynomial relationship. Results show that PLCCs to MOS varying between 0.67 and 0.88 can be achieved, depending on the considered content. In a subsequent study [6], the same authors developed a FR metric called *BitDance* by extracting and comparing different color and geometry statistics from both the original and

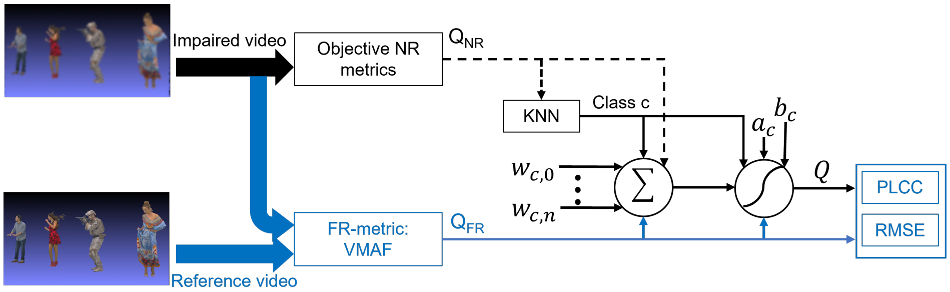

Fig. 1: Block diagram of the presented methodology. The parts in black are used at client side for real-time quality assessment. The blue parts are added at server side for training and evaluation purposes.

the distorted content. They realised comparable or even improved results in comparison with well-known metrics such as *PCQM* and *PointSSIM-Color*. Viola *et al.* created a RR quality metric by extracting color statistics and constructing histograms and correlograms from both the original and the distorted sequence [7]. The distance between both is used to predict the subjective MOS by applying a curve-fitting approach. Following up on this work, the authors created a second RR metric based on a weighted combination of feature differences in terms of geometry, luminance and normal [8]. Evaluating both metrics, PLCCs up to 0.90 for the subjective MOS are achieved for a single publicly available dataset. In our own previous work [9], we presented an objective and subjective quality evaluation of point cloud streaming for multiple scenarios in terms of bandwidth, rate adaptation, viewport prediction and user motion. The results show high correlation with MOS for traditional video metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Video Quality Metric (VQM). We further indicated that the subjective perception of volumetric media lays within a very small interval of the total range of the objective metrics, which might be a result of the inclusion of (too much) background during the quality metric calculation. In our second study [2], a thorough correlation analysis of both FR and NR objective metrics to subjective Double Stimulus (DS) and Single Stimulus (SS) MOS for different volumetric streaming scenarios was performed. Correlations were obtained between each of the NR metrics, the SS and the DS MOS and Video Multimethod Assessment Fusion (VMAF). It was noticed that VMAF correlates very well to both SS and DS MOS, with PLCC values above 0.92 overall and even above 0.94 on a per-video basis. In terms of NR metrics, the blur ratio (BRT) showed the strongest correlation to both VMAF and the subjective benchmarks of all NR features in all three videos in terms of PLCC.

The state of the art has shown that FR metrics such as SSIM [10], VQM [11] and VMAF [12] have the highest accuracy, where VMAF has obtained the best results. However, they require high computational complexity as well as the simultaneous access of both the original and distorted content. These circumstances make them unsuitable for real-time evaluations. NR metrics such as blur, blockiness and noise, however, are more computationally friendly but lack in straightforward correlation towards subjective perception, as we showed in our previous work [2]. Encouraged by our previous results, in this paper, we present a novel NR objective metric for real-time quality assessment of point cloud streamed content, which leverages the correlation patterns of individual NR metrics. By means of a K-Nearest Neighbours (KNN) clustering mechanism and a per-cluster sigmoidal mapping of a linear combination of NR pixel-based features, the model is able to predict the user's perceived quality with high correlation to objective FR benchmarks (VMAF). The model is evaluated on a broad, objectively labeled (VMAF) point cloud dataset.

The remainder of this paper is structured as follows. Section 2 presents our modelling approach. Section 3 discusses the dataset used, as well as the results. Finally, Section 4 concludes the paper.

2. METHODOLOGY

The purpose of this work is to create a content-independent, computationally friendly, and accurate objective NR model for quality assessment of point cloud video. To this end, the approach presented in Figure 1 was followed. This method calculates the perceived quality as a sigmoidal fitting of a Linear Regression (LR) of NR metrics towards an objective FR benchmark, where the different weights and parameters are determined based on the particular class of the video. This classification is needed as our previous work has shown that different types of videos can rely on totally different types of NR metrics for quality estimation [3].

Whenever a new video is received at the client side (indicated in black), a set of NR-metrics Q_{NR} (i.e., bandwidth (BW), blur (BLU) [13], blur ratio (BRT) [13], noise (NOI) [13], noise ratio (NRT) [13], blockiness (BLK) [14] and Spatial Information (SI) [15]) are calculated/extracted on each of the incoming frames. Once a sufficiently large portion of the video (i.e., a couple of seconds) is received, the obtained metrics are averaged to characterize the given video (apart from SI, where the maximum is taken by definition [15]). The obtained characterization is fed to a KNN classifier, which is chosen for its fast training and evaluation times as well as its intuitive interpretation. This classifier is pre-trained at server side based on prior available sequences, in order to obtain the class c of the given video. Note that KNN also allows for easy updating of this classification whenever new content is added to the server-side database. Based on this class, the weights $w_{c,i}$ for a linear combination of the NR-metrics x_i as well as the parameters a_c and b_c of a sigmoidal mapping (Equation 1) are determined, where the latter is based upon the well-known Quality of Service (QoS)-QoE relationship proposed by Fiedler et al. [16] and previously applied in our former work [3]. The subsequent calculation of both, results in the quality prediction Q.

$$Q = \sigma\left(w_0 + \sum_{i=1}^{7} w_i \cdot x_i\right) \text{ with } \sigma(x) = \frac{1}{1+e^{ax-b}} \quad (1)$$

At server side (indicated in blue), both the LR and the sigmoid are trained by minimizing Mean Squared Error (MSE) against a FR benchmark Q_{FR} which is known to correlate strongly to subjective scores (e.g., VMAF [2]). Note that this metric cannot be calculated at client side as the undistorted content is unavailable. Furthermore, Q_{FR} is also used for evaluation of the obtained models (e.g., in terms of PLCC and Root Mean Squared Error (RMSE)). Note that the calculation of Q_{FR} on new content is only needed if it fundamentally differs from the current dataset. This can be done on the server side, however, where computational and time-related requirements are less stringent. By sending the appropriate weights and parameters as well as the classifier to the client at video request, this provides all tools for client-side quality estimation.

3. EVALUATION

In this Section, the dataset used to evaluate our algorithm will be first described, followed by an analysis of our method.

3.1. Dataset

For the evaluation of the model, we took the adaptive streaming point cloud dataset from our previous work [2,17] and extended it with more conditions and video scenes. As a result, we obtain a set of sixteen source videos (Table 1) between 18 and 50 seconds of length. Each sequence contains the generated viewport of a scene consisting of four point cloud objects from the 8i dataset [18], each with a different setup of the figures (circle, line, semicircle, square) and camera movement (rotation, zoom, pan, zigzag). These objects were encoded using the V-PCC encoder [19] with five reference quality representations, each between 2.4 Mb/s and 53.5 Mb/s. Afterwards, the resulting videos were streamed using the Dynamic Adaptive Streaming over HTTP (DASH) protocol with multiple combinations of bandwidth (15, 20, 60, 100, 140 and ∞ Mb/s), resolutions (800 × 592 and 1920 × 1080), buffer lengths (0, 1, 2, 3 and 4 seconds) and allocation algorithm (greedy, hybrid and uniform). The frame rate was fixed at 30 Frames Per Second (FPS). As such, a total of 453 sequences was obtained. For each of these, the seven objective NR metrics mentioned in Section 2 were calculated in addition to the FR VMAF, which will be used as benchmark. In addition, videos 1-3 were also subjectively annotated with both DS and SS MOS based on two subjective experiments with 30 subjects each. Note that in our previous work [2], we already showed the overall high correlations of VMAF to SS and DS MOS for projected point cloud QoE. Therefore, there was decided to evaluate against VMAF to illustrate the envisioned system without subjective scoring, as the latter is not scalable and very costly in terms of time and effort and therefore not suited for live-streaming environments.

Table 1: Summary of the 16 videos in the dataset in terms of point cloud constellation, duration and camera movement.

Video	Setup	Dur.	Camera movement
1	line	24s	pan left-to-right and back (angle)
2	semi-circle	18s	zoom-in/zoom-out + rotate to next (object 1 and 2)
3	semi-circle	18s	zoom-in/zoom-out + rotate to next (object 3 and 4)
4	circle	24s	outside rotation
5	circle	24s	outside rotation + zoom in/zoom out
6	circle	24s	outside rotation + zoom in/zoom out in between figures
7	line	24s	pan left-to-right and back (angle)
8	line	24s	pan left-to-right and back (frontal)
9	line	24s	rotate left-to-right and back
10	semi-circle	50s	zoom-in/zoom-out + rotate to next
11	semi-circle	24s	rotate left-to-right and back
12	semi-circle	24s	rotate + pause on figure
13	square	24s	outside rotation
14	square	24s	outside rotation + zoom-in/zoom-out
15	square	24s	outside rotation + zoom-in/zoom-out
16	square	24s	zig-zag

3.2. Results

We analysed the performance of the quality method in two stages: (i) the KNN classifier and the (ii) per-class quality modelling as a combination of a LR and a sigmoidal mapping. First, the KNN classifier was implemented using *Python's SciKit Learn* library [20], using 10 runs of the algorithm with a maximum of 300 iterations per run. The relative tolerance is set to 0.0001. Note that the NR metrics that do not lay

Table 2: The four different clusters obtained by the KNN algorithm, as well as their average PLCC and RMSE per fold for each cluster.

Cluster	Videos	PLCC	RMSE
0	2, 7-9, 11, 12, 14-16	0.983	0.080
1	4-6, 13	0.994	0.047
2	1, 3	0.985	0.135
3	10	0.841	0.053

Table 3: Average PLCCs and RMSEs per video of the proposed solution, compared with the cases with and without clustering and sigmoidal mapping.

(a) PLCC

	Without clustering	**With clustering**
LR	0.826	0.942
LR + sig. map.	0.869	**0.977**

(b) RMSE

	Without clustering	**With clustering**
LR	0.114	0.095
LR + sig. map.	0.199	**0.077**

within the $[0, 1]$ interval by construction (BW, BLU, NOI, SI) are first normalized using min-max scaling as the KNN is a distance-based clustering mechanism. The number of clusters is chosen by optimizing the so-called *Silhouette Coefficient* [21]. On the given dataset, this results in a total number of four clusters with a Silhouette Coefficient of 0.606. To allow for easy Cross Validation (CV) within each cluster, videos that are divided over multiple clusters are completely assigned to the cluster with the highest number of samples. This has resulted in the clusters shown in Table 2. Within each cluster, multiple iterations of the modelling approach were conducted where in each run all configurations of one of the videos were held out as a test set while the others were used for training. For cluster 3, which only contains one video, a 5-fold CV was performed on the 33 configurations of video 10 itself. Note that the normalization parameters for BLU, NOI, SI and BW were recalculated each iteration on the training set only to avoid data leakage.

Table 2 shows the average PLCC and RMSE obtained per test fold within each cluster. Note that, for clarity and conciseness, the training scores were omitted from the manuscript. They show similar results as the test scores, however, such that overfitting is unlikely. As can be seen, high correlations were obtained (>0.98), with the exception of cluster 3 where the PLCC is limited to 0.84. A possible explanation is the limited amount of data, such that not all possible configurations are seen per training iteration. Given the fact that it ends up in an isolated cluster, however, can also mean that the video consists of fundamentally different characteristics in comparison with the other videos, thus proving to be an outlier in the dataset. This is, for example, illustrated by the significant difference in blur ratio as is shown in the noise ratio vs. blur ratio plot in Figure 2. Furthermore, a higher RMSE of cluster 2 can be noticed, while still resulting in a high PLCC of 0.985. This is a result from the fact that the VMAF of video 3 is consistently underpredicted when trained on video 1 and vice versa. As such, it can be expected that additional data would provide a more reliable presentation of the content within the class or possibly an even improved clustering mechanism. In Table 3, the average obtained PLCCs and RMSEs per video can be found compared to the cases with and without clustering and sigmoidal mapping. As can be seen, the clustering mechanism prior to modelling is improving the non-clustering case by 0.108-0.116 in terms of PLCC and 0.019-0.122 in terms of RMSE. Moreover, the sigmoidal mapping shows to be beneficial with PLCC and RMSE gains of 0.035-0.043. In terms of RMSE, however, the sigmoidal mapping only shows to be beneficial when combined with clustering (0.028 gain). Without clustering, the excess of data seems to withhold the model from calculating an appropriate fit.

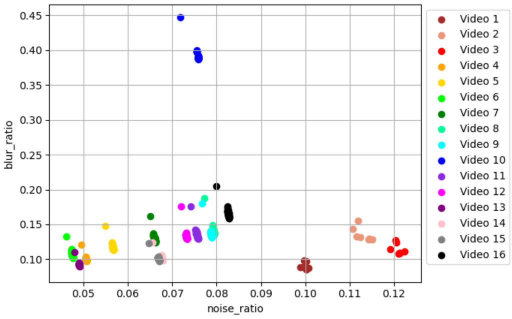

Fig. 2: Scatterplot showing the noise ratio and blur ratio of the different videos.

4. CONCLUSION

In this paper, we have presented an NR QoE assessment model consisting of a KNN-based clustering mechanism followed by a per-cluster LR and sigmoidal mapping. PLCCs up to 0.977 and RMSEs down to 0.077 towards VMAF are achieved. This method clearly shows to outperform the cases where either the mapping or the clustering is omitted, therefore showing its potential. As QoE modelling is performed at video level in this work, it is worth further exploring whether these findings still hold on a per-Group Of Pictures (GOP) or even a per-frame level. Furthermore, the accuracy of the proposed metric towards other types of compression distortion such as G-PCC, should be researched.

5. ACKNOWLEDGEMENTS

Sam Van Damme, Jeroen van der Hooft and Maria Torres Vega are funded by the Research Foundation Flanders (FWO), grant numbers 1SB1822N, 1281021N and 12W4819N, respectively.

6. REFERENCES

[1] MPEG, "MPEG 3DG and Requirements - Call for Proposals for Point Cloud Compression V2," `https://bit.ly/2RSWdWe`, 2017.

[2] S. Van Damme, M. Torres Vega, and F. De Turck, "A Full- and No-Reference Metrics Accuracy Analysis for Volumetric Media Streaming," in *International Conference on Quality of Multimedia Experience*, 2021.

[3] S. Van Damme, M. Torres Vega, J. Heyse, F. De Backere, and F. De Turck, "A Low-Complexity Psychometric Curve-Fitting Approach for the Objective Quality Assessment of Streamed Game Videos," *Signal Processing: Image Communication*, vol. 88, 2020.

[4] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun, "Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration," *IEEE Transactions on Multimedia*, 2020.

[5] R. Diniz, P. G. Freitas, and M. C. Q. Farias, "Towards a Point Cloud Quality Assessment Model using Local Binary Patterns," in *International Conference on Quality of Multimedia Experience*, 2020.

[6] Rafael Diniz, Pedro Garcia Freitas, and Mylène C. Q. Farias, "Color and geometry texture descriptors for point-cloud quality assessment," *IEEE Signal Processing Letters*, vol. 28, pp. 1150–1154, 2021.

[7] I. Viola, S. Subramanyam, and P. Cesar, "A Color-Based Objective Quality Metric for Point Cloud Contents," in *International Conference on Quality of Multimedia Experience*, 2020.

[8] I. Viola and P. Cesar, "A Reduced Reference Metric for Visual Quality Evaluation of Point Cloud Contents," *IEEE Signal Processing Letters*, vol. 27, 2020.

[9] J. van der Hooft, M. Torres Vega, C. Timmerer, A. C. Begen, F. De Turck, and R. Schatz, "Objective and Subjective QoE Evaluation for Adaptive Point Cloud Streaming," in *International Conference on Quality of Multimedia Experience*, 2020.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004.

[11] M. H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, 2004.

[12] A. Aaron, Z. Li, M. Manohara, J. Y. Lin, E. C. Wu, and C. Kuo, "Challenges in Cloud Based Ingest and Encoding for High Quality Streaming Media," in *IEEE International Conference on Image Processing*, 2015.

[13] Min Goo Choi, Jung Hoon Jung, and Jae Wook Jeon, "No-reference image quality assessment using blur and noise," *International Journal of Computer Science and Engineering*, vol. 3, no. 2, pp. 76–80, 2009.

[14] Cristian Perra, "A low computational complexity blockiness estimation based on spatial analysis," in *2014 22nd Telecommunications Forum Telfor (TELFOR)*, 2014, pp. 1130–1133.

[15] Pradip Paudyal, Federica Battisti, and Marco Carli, "Impact of video content and transmission impairments on Quality of Experience," *Multimedia Tools and Applications*, vol. 75, no. 23, pp. 16461–16485, Dec 2016.

[16] Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[17] J. van der Hooft, M. Torres Vega, T. Wauters, C. Timmerer, A. C. Begen, F. De Turck, and R. Schatz, "From Capturing to Rendering: Volumetric Media Delivery with Six Degrees of Freedom," *IEEE Communications Magazine*, vol. 58, no. 10, 2020.

[18] E. d'Eon, T. Myers, B. Harrison, and P. A. Chou, "Joint MPEG/JPEG Input. 8i Voxelized Full Bodies - A Voxelized Point Cloud Dataset," `https://jpeg.org/plenodb/pc/8ilabs/`, 2017.

[19] "V-PCC," `https://github.com/MPEGGroup/mpeg-pcc-tmc2`.

[20] "SciKit Learn," `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`.

[21] "Silhouette Score," `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`.