TOWARDS JOINT FRAME-LEVEL AND MOS QUALITY PREDICTIONS WITH LOW-COMPLEXITY OBJECTIVE MODELS

Joel Jung, Alexandre Giraud, Meijia Song, Songnan Li, Xiang Li, Shan Liu

Tencent Media Lab, Palo Alto

ABSTRACT

The evaluation of the quality of gaming content, with low-complexity and low-delay approaches is a major challenge raised by the emerging gaming video streaming and cloud-gaming services. Considering two existing and a newly created gaming databases this paper confirms that some low-complexity metrics match well with subjective scores when considering usual correlation indicators. It is however argued such a result is insufficient: gaming content suffers from sudden large quality drops that these indicators do not capture. In addition to proposing three new low-complexity models based on various machine learning techniques, this paper introduces a new indicator to capture sudden quality variations and reports poor results for most of the models when applying this indicator. Consequently, an original way to train the models, using jointly the subjective scores and the frame level scores of a full-reference metric, is proposed. The high correlation through traditional indicators is preserved, while the efficiency on the new indicator is drastically improved.

Index Terms— *objective quality metric, cloud-gaming, video quality evaluation, low-complexity*

1. INTRODUCTION

The streaming of video content is growing rapidly through new emerging services. Passive viewing of gaming content using famous platforms gathers hundreds of millions of viewers per year. Cloud-gaming services are deploying extremely fast thanks to 5G: the player sends actions to the cloud where the scene is rendered and 2D encoded, to be sent back to PCs or mobile devices. This makes gaming accessible to low-end client devices because heavy processes of rendering the game is done in the cloud.

Measuring the quality is fundamental for gaming services, and in particular for cloud gaming, to control encoders and ensure customers' satisfaction. Subjective tests are the ultimate and most reliable solution. Unfortunately, they are time consuming and costly. Measuring the quality in an automatic way is consequently needed. In this work, we focus on low-complexity and low-delay metrics for cloud-gaming services, in passive scenarios, without considering interaction quality and network degradations.

Not so many databases with subjective scores dedicated to gaming are available. A first contribution of this paper is to produce a new database of 170 scenes with scores obtained by crowdsourcing from 64 workers. Several objective models already exist in the literature: a second contribution of the paper is to evaluate state-of-the-art metrics on an extended dataset of gaming content, of different nature, obtained from different subjective test methodologies and with different encoders. Low-complexity and low-delay are two major aspects of quality measurements in the cloud or in a client device. As another contribution, three new low-complexity machine learning based models are proposed and evaluated.

Gaming content has well known characteristics such as extremely fast motion and specific motion and texture patterns. We argue it also contains sudden scene rotations (especially in shooting games) and severe illumination changes (resulting from explosions for instance) that undermine video encoders' efficiency, yielding severe sudden quality variations that are currently not considered during the design of objective quality models. A fourth contribution of the paper is to propose an indicator able to capture how much an objective metric reflects sudden quality changes. This indicator, tested on state-of-the-art as well as three new proposed models, reports insufficient results. The last contribution of the paper is consequently to introduce a modification of the training methodology to improve the models' performance according to this indicator, without affecting the efficiency according to usual correlation indicators.

The remainder of the paper is organized as follows: section 2 describes the dataset used for the proposed evaluations and developments. Section 3 lists state-of-the-art metrics for gaming content and evaluates some of them under practical gaming conditions. Section 4 proposes and evaluates three new machine learning based models. In Section 5, a new indicator is suggested to better handle gaming content specificities and metrics are evaluated accordingly. Section 6 proposes an improvement of the models to jointly optimize the proposed indicator and classical ones. Section 7 concludes the paper.

2. GAMING CONTENT DATASETS

Datasets dedicated to gaming content are rare. To make the database more complete in terms of content, variety of bitrate ranges and coding conditions, we have created a new dataset, called Tencent Gaming Data Set (TGDS), and merged it with two public datasets.

2.1. KUGVD and CGVDS datasets

From the Kingston University Gaming Video Dataset (KUGVD) [1], we use 30 Processed Video Segments (PVS) extracted from 6 games. The PVS are 1080p@30fps of 30 s duration, encoded with H.264 (with I, P and B frames), at five bitrates from 600 kbps to 4 Mbps. From the Cloud-Gaming Video Data Set (CGVDS) [2], we use 39 PVS extracted from 13 games. The PVS are 1080p@60fps of 30 s duration, encoded with the hardware accelerated implementation of H.264 (NVENC) [3] with the low-latency high quality profile, at three bitrates from 2 Mbps to 6 Mbps. For both datasets, the subjective evaluation has been performed in a laboratory, using an Absolute Category Ranking (ACR) scale (simple or extended) with 17 and 20 subjects respectively.

2.2. New TGDS dataset

The new dataset contains 170 PVS, from 34 different scenes extracted from five famous games (Fortnite, Blade&Soul, Path of Exile, League of Legends, The Witcher). The PVS are 1080p@60fps of 10 s duration, encoded with a proprietary H.264 encoder, at five bitrates from 6 Mbps to 30 Mbps. A low-delay configuration, typical for cloud-gaming, was used. The subjective tests were performed through crowdsourcing, mixing recommendations provided in P.809 [4] for subjective evaluation methods for gaming content and in P.808 [5] for guidelines on how to perform valid crowdsourcing tests for speech. A five-grades ACR scale has been used and 64 workers have scored the PVS. The scores obtained from crowdsourcing are known to be less reliable than those from laboratory tests: some workers might be less focused, have inappropriate viewing conditions, *etc*. However, thanks to the larger amount of workers, it is possible to apply strict rules to discard wrong scores. As far as we know, there is no recommendation specifically related to crowdsourcing database analysis, to reject scores or workers. We have decided to apply the following rules, consecutively:

1. Rate point inconsistency: considering five different rate points, with RP1 and RP5 being respectively the highest quality and lowest quality, the scores of a worker for a sequence are rejected when RP1 MOS is below RP5 MOS.
2. P.913 recommendation [6]: P.913 adjusts the scores of each viewer, by removing bias, when compared to other viewers.
3. BT.500 [7] recommendation: an analysis of each worker's scores distribution allows BT.500 to detect a worker scoring too far from others. If the situation occurs often and if the outlier scores are on the lower and higher end of the scale, the worker is discarded.

The rate point inconsistency method has discarded 8.5% of the scores. This high value highlights how important it is to properly post-process a database obtained with a crowdsourcing approach. After this step, BT.500 has further discarded 25 workers: the number of remaining scores is 6041. As a result of this dataset post-processing, an excellent confidence interval of 0.33 is achieved.

3. RELATED WORK AND EVALUATION OF STATE-OF-THE-ART METRICS

In 2018, a first evaluation of quality metrics on gaming content was presented in [7]. While good correlation was reported between MOS and VMAF, poorer results were achieved by no-reference metrics. Progressively, new methods were proposed, significantly improving the situation. In [9], NR-GVQM no-reference machine learning-based metric is proposed, trained on low-level image features. In [10], another no-reference pixel-based metric is proposed. In [11], two other metrics are proposed: NR-GVSQI and NR-GVSQE, using supervised learning algorithms trained either on MOS or on VMAF at the frame level. Unfortunately, all these models achieve great match of either the MOS on a few seconds or VMAF scores at frame level, but without addressing one major issue related to gaming content specificity, raised in section 5.

In this paper, seven other state-of-the-art metrics are evaluated. First, three full reference ones are considered: Peak Signal to Noise Ratio (PSNR) fidelity metric, Structural Similarity Index Metric (SSIM), and Video Multi-Method Assessment Fusion (VMAF) [11]. To evaluate how more complex approaches with access to decoded pixels, but without reference can perform, two deep learning based methods are selected. Deep Bilinear Convolutional Neural Network (DBCNN) [12] is based on two deep convolutional neural networks: S-CNN trained on the Waterloo database, VGG-16 trained on ImageNet, both handling different kind of distortions. The features from the two CNNs are pooled bilinearly. NDNetGaming [13] is another CNN-based metric developed specifically for gaming content. It uses transfer learning: it is first trained on VMAF scores, then the last layers are retrained on a smaller dataset containing MOS from human observers. In the low-complexity category, two ITU-T bitstream-based models are selected. They are no-reference metrics, with no access to decoded pixels. P.1203.1 [15] predicts the visual quality in the context of adaptive and progressive-download-type media streaming. We use its mode 3 that has access to the full bitstream: all QPs are parsed. The following equation is the core of the method, where *quant* reflects the quantization strength, and the q_i are coefficients: $MOS_q = q_1 + q_2 * exp(q_3 * quant)$

Its successor P.1204.3 [16], includes a parametric part, close to P.1203.1, and a machine learning part. The latter is a Random Forest that predicts the residual between the parametric model and the quality score of the video. A linear combination of the sub-models is processed to supply the final score.

The three databases, KUVGD, CGVSD, TGDS are merged

	PSNR	SSIM	VMAF	P.1203	P.1204	DBCNN	NDNet
RMSE	0.58	0.56	0.44	0.47	0.46	0.50	0.42
PLCC	0.67	0.68	0.81	0.79	0.80	0.76	0.83
SRCC	0.65	0.78	0.82	0.80	0.80	0.74	0.82

Table 1: state-of-the-art metrics evaluated on test set.

	VQMCG.a	VQMCG.b	VQMCG.c
RMSE	0.40	0.32	0.29
PLCC	0.87	0.91	0.93
SRCC	0.88	0.91	0.92

Table 2: performance of new low-complexity no-reference metrics.

to form a dataset of 239 PVS and 53 different scenes. The dataset is randomly split into two subsets, one for the training (186 PVS) and one for the testing (53 PVS). Experimental results are reported in Table 1 for the different objective metrics, according to three indicators: the Root Mean Square Error (RMSE), the Pearson Linear Correlation Coefficient (PLCC), and the Spearman Ranked Order Correlation Coefficient (SRCC).

The dataset mixes PVS produced by different codecs with MOS obtained under different test methodologies. To compensate over- or under-estimation of quality due to these varying parameters, a linear mapping of the scores is applied as suggested in P.1401 [16], where linear coefficients are learnt on the training set. The PSNR and SSIM metrics provide insufficient correlation, yet the performance of the other metrics is good, both in terms of error and correlation. DBCNN is slightly below others, most likely because it has not been trained on gaming content. Interestingly, the bitstream-based models P.1203.1 and P.1204.3 are very efficient at predicting the MOS even though they are low-complexity models, confirming results presented in [17]. VMAF remains slightly better, and NDNetGaming, no-reference but trained on gaming content, shows the best performance.

4. NEW LOW-COMPLEXITY MODELS

We propose three new models and evaluate them under the test conditions described in section 3. The three metrics are low-complexity, no-reference, and do not require decoded pixels. They rely on features extracted from the parsed bitstream, including the *quant* feature used in P.1203.1, augmented by 8 other features: frame size, bitrate, spatial complexity, and average number of: P macroblocks per frame, 8x8 blocks, blocks with 8x8 transform size, blocks without frequential transform, skipped blocks. The metrics denoted VQMCG for Video Quality Metrics for Cloud Gaming, provide a quality score on a 5-level ACR scale. VQMCG.a is a weighted linear combination of the features. The weight for each feature is learnt on the training set with a gradient descent. VQMCG.b applies Support Vector Regression (SVR). It is a supervised learning algorithm that tries to map the features space with the video quality by finding a hyperplane that encompasses most of the data point from the training set. The coefficients of this hyperplane are obtained using the Lagrange multiplier technique, and its equation is used to make the predictions. VQMCG.c uses a Multi-Layer Perceptron (MLP). It is a neural network with an architecture consisting of a sequence of fully connected layers where the neurons are linked by weights. During the training, the weights are initialized with Glorot initialization and trained with the back-propagation algorithm using the Adam optimizer. The MLP has four layers with 100, 50, 25 and 10 neurons respectively, activated by a ReLU function.

The performance of the proposed models is provided in Table 2. While better than VMAF and NDNetGaming (see Table 1), VQMCG.a is the least efficient proposed models. VQMCG.b and VQMCG.c both outperform existing methods on all indicators. Although the three variants remain of low complexity, the efficiency increases with the complexity, from VQMCG.a to VQMCG.c.

Generally, we observe high performance for the learning-based models (P.120x, NDNetGaming and VQMCG.x), in terms of RMSE and PLCC. This might appear surprising at a first sight. Our deeper analysis reveals that games are made of repetitive visual characteristics, as being computer generated content obtained from a game engine. The motion pattern, the color diversity, the backgrounds are among such characteristics. Similar scenes are often repeated in games, offering spatial and temporal similarities, explaining why gaming content is well adapted to learning-based models.

5. NEW INDICATOR FOR GAMING

As seen in previous sections, satisfying results are achieved by several metrics, including low-complexity ones, and by the new proposed models, when matching a MOS on a video segment of a few seconds. By analyzing the temporal evolution of the quality of several famous video games, we noticed large and abrupt quality changes. They result from specific characteristics of gaming contents, among which sudden and fast rotations of a character, or explosions. These events generate high temporal and spatial changes difficult to handle by codecs, especially under low-delay and strict bitrate constraints. The motion estimation is challenged and the rate-control algorithms increase the number of Intra blocks. The low-complexity models evaluated in section 3 and 4 have difficulties to reflect these sudden changes of quality. This is understandable because they were not designed with this specific goal in mind. Figure 1 reflects the temporal evolution of objective quality scores for the BlackDesert game. The orange curve is a low-complexity bitstream-based model. On the complete test set, according to RMSE and PLCC indicators, it performs similarly as VMAF. When looking at a finer temporal granularity, it is however obvious that VMAF performs much better.

To quantify and handle this issue a new indicator is proposed: the Frame Variation Match (FVM). It measures the ability of a model to reflect large and sudden quality variations when a reference metric reports similar kind of

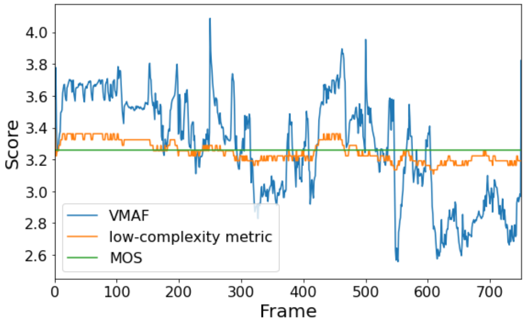

Figure 1: inability of low-complexity models to reflect quality changes.

P.1203.1 / #	P.1204.3 / #	VQMCG.a	VQMCG.b	VQMCG.c
2% / 36%	4% / 29%	71%	24%	31%

Table 3: value of the indicator for the low-complexity models.

variation. FVM, as depicted in Figure 2, counts the percentage of time when the model has a quality change *varMod* above a threshold *th*, when the reference metric also has a quality change *varRef* above *th*, and in the same direction, in the same window *W*.

While any full-reference metric could be considered, in this experiment VMAF is selected as the reference because it is widely used, and sufficiently reliable at frame level. Even if it is not a perfect reference, when designing no-reference low-complexity models any step towards matching such a reference at the frame level is of interest. Table 3 reports the FVM of the low-complexity models, with *th=1* (score), *W=10* (frames). Results for NDNetGaming and DBCNN are not provided on purpose: comparison with a VMAF reference does not make sense for metrics that are possibly as reliable as VMAF. Said differently it would have been acceptable to use them as a reference, instead of VMAF. We observe that the original bitstream-based models P.1203.1 and P.1204.3 fail to reflect the quality changes. We have updated the two standards to provide frame level results (marked with a # in Table 3). This modification significantly improves the FVM, that becomes in line with VQMCG.b and VQMCG.c. Only VQMCG.a reports a much higher FVM: a deeper analysis shows that this model tends to over-estimate the quality peaks in some cases, which artificially increases FVM value.

6. NEW TRAINING METHOD FOR GAMING

A new training method is applied to improve the accuracy of the VQMCG models on the FVM indicator. Some existing models were trained using both MOS and objective scores. For instance, NDNetGaming applies transfer-learning, learning first from VMAF scores, and further refining the last layers on MOS. The approach we propose, using the two information jointly, is drastically different. The objective scores are used at a frame level, and the MOS just serves as an offset applied to the objective scores: the models are trained on VMAF scores shifted by a constant value, that centers VMAF scores on the MOS. Using this new reference for the training, it is possible to improve the performance of FVM vs VMAF, without degrading the other indicators vs MOS. Table 4 shows the results for the proposed models with the new training process (marked *). The performance on the FVM indicator is increased for all models (see Table 3), except for VQMCG.a (related to the explanation given in section 5). The performance on usual indicators has only slightly decreased (see Table 2). Still VQMCG.a* performance remains slightly better than NDNetGaming, and VQMCG.b* and VQMCG.c* much better (see Table 1).

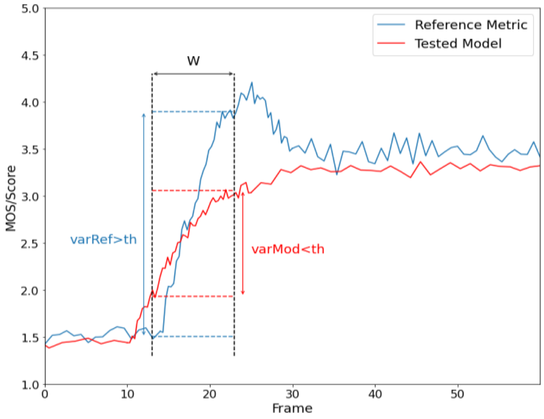

Figure 2: frame Variation Match (FVM) indicator.

	VQMCG.a*	VQMCG.b*	VQMCG.c*
RMSE / MOS	0.40	0.26	0.30
PLCC / MOS	0.86	0.94	0.92
SRCC / MOS	0.87	0.93	0.91
FVM / VMAF	36%	51%	50%

Table 4: correlation of the new models with the new training method.

7. CONCLUSION

This paper addresses the objective quality evaluation of gaming content with low-complexity approaches. We create a new gaming dataset using crowdsourcing methodology and merge it with other datasets to evaluate seven state-of-the-art metrics of different nature. We additionally propose three new low-complexity models, with different machine learning techniques. We report satisfactory results on the usual correlation indicators versus MOS for the state-of-the-art methods and excellent ones for the proposed models. We however argue that this performance is insufficient for gaming contents with sudden large quality drops that these indicators cannot capture. To tackle this problem, we first propose a new indicator able to reflect such quality changes. When assessing the models with this indicator, unsatisfactory results are obtained. We consequently propose an original process of training models: the objective scores are shifted towards the MOS before the training. The efficiency on traditional indicators is preserved, while the performance on the new indicator is drastically improved.

8. REFERENCES

[1] N. Barman, E. Jammeh, S. A. Ghorashi and M. G. Martini, "No-Reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications," in *IEEE Access*, vol. 7, pp. 74511-74527, 2019, doi: 10.1109/ACCESS.2019.2920477.

[2] S. Zadtootaghaj, S. Schmidt, S. Shafiee Sabet, S. Möller, and C. Griwodz, “Quality estimation models for gaming video streaming services using perceptual video quality dimensions”, Proceedings of the 11th ACM Multimedia Systems Conference. Association for Computing Machinery, New York, NY, USA, 213–224.

[3] R. Arzumanyan, “Turing H.264 Video Encoding Speed and Quality. NVIDIA Developer Blog”, Retrieved August 2021 from https://devblogs.nvidia.com/turing-h264-video-encoding-speed-and-quality/.

[4] ITU-T Recommendation P.809. 2018. Subjective evaluation methods for gaming quality. Geneva Switz. Int. Telecommun. Union.

[5] ITU-T Recommendation P.808. 2021. Subjective evaluation of speech quality with a crowdsourcing approach. Geneva Switz. Int. Telecommun. Union.

[6] ITU-T Recommendation P.913. 2021. Subjective evaluation of speech quality with a crowdsourcing approach. Geneva Switz. Int. Telecommun. Union.

[7] ITU-T Recommendation BT-500-14. 2019. Methodologies for the subjective assessment of the quality of television images. Geneva Switz. Int. Telecommun. Union.

[8] N. Barman, S. Schmidt, S. Zadtootaghaj, M.G. Martini, S. Möller, “An evaluation of video quality assessment metrics for passive gaming video streaming”, in: Proceedings of the 23rd packet video workshop. ACM, pp 7–12, 2018.

[9] S. Zadtootaghaj, N. Barman, S. Schmidt, M.G. Martini, S. Möller, “NR-GVQM: a no reference gaming video quality metric”, in: IEEE international symposium on multimedia (ISM). IEEE, pp 131–134, 2018.

[10] S. Göring, R.R.R. Rao, A. Raake, “Nofu- a lightweight no-reference pixel based video quality model for gaming QoE”, in: 11th international workshop on quality of multimedia experience (QoMEX), pp 1–6, 2019.

[11] N. Barman, E. Jammeh, S. A. Ghorashi and M. G. Martini, "No-Reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications," in *IEEE Access*, vol. 7, pp. 74511-74527, 2019, doi: 10.1109/ACCESS.2019.2920477.

[12] Netflix. 2020. Perceptual video quality assessment based on multi-method fusion. Retrieved June, 2021 from https://github.com/Netflix/vmaf.

[13] W. Zhang, K. Ma, J. Yan, D. Deng and Z. Wang, "Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 1, pp. 36-47, Jan. 2020, doi: 10.1109/TCSVT.2018.2886771.

[14] M. Utke, S. Zadtootaghaj, S. Schmidt, S. Bosse, S. Möller, “NDNetGaming - development of a no-reference deep CNN for gaming video quality prediction”, Multimed Tools App, 2020. Doi: 10.1007/s11042-020-09144-6.

[15] ITU-T Recommendation P.1203. 2017. Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport. Geneva Switz. Int. Telecommun. Union.

[16] ITU-T Recommendation P.1204. 2020. Video quality assessment of streaming services over reliable transport for resolutions up to 4K. Geneva Switz. Int. Telecommun. Union.

[17] R.R.R. Rao, S. Göring, R. Steger, S. Zadtootaghaj, N. Barman, S. Fremerey, S. Moller, A. Raake, "A Large-scale Evaluation of the bitstream-based video-quality model ITU-T P.1204.3 on Gaming Content," *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1-6, doi: 10.1109/MMSP48831.2020.9287055.

[18] ITU-T Recommendation P.1401. 2020. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. Geneva Switz. Int. Telecommun. Union.