SPECIALISED VIDEO QUALITY MODEL FOR ENHANCED USER GENERATED CONTENT (UGC) WITH SPECIAL EFFECTS

Anne-Flore Perrin, Yejing Xie, Tao Zhang, Yiting Liao, Junlin Li, Patrick Le Callet

SPECIALISED VIDEO QUALITY MODEL FOR ENHANCED USER GENERATED CONTENT (UGC) WITH SPECIAL EFFECTS

Anne-Flore Perrin[§] *Yejing Xie*[§] *Tao Zhang*[†] *Yiting Liao*[†] *Junlin Li*[†] *Patrick Le Callet*[⋆]

[§] Nantes Université, Ecole Centrale Nantes, CAPACITES SAS, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[†] ByteDance Inc., Mountain View, CA, US
[⋆] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-4000 Nantes, France

ABSTRACT

User Generated Content (UGC) refers to media generated by users for end-consumers that represent most of the media exchange on social media. UGC is subject to acquisition and transmission limitations that disable access to the pristine, i.e., perfect source content. Evaluating their quality, especially with current pre- and post-processing algorithms or filters, is a major issue for most off-the-shelf full-reference quality metrics. We propose to conduct a benchmark on existing full-reference, non-reference, and aesthetic quality metrics for UGC with special effects. We aim to identify the challenges posed by both UGC and filtering. We then propose a new combination of metrics tailored to enhanced and filtered UGC, which reaches a trade-off between complexity and accuracy.

***Index Terms*—** Video Quality Metric, Fine-Tuning, User Generated Content, Enhancement filters, Crowdsourcing

1. INTRODUCTION

With the rapid development of social media and video sharing platforms, UGC has a significant and rising share of internet traffic. As existing encoding-quality assessment systems are designed for pristine originals, these systems commonly fail when forced to consider original content containing distortions - a common characteristic of UGC. To design, improve, or select efficient compression recipes, robust quality metrics tailored to UGC are inevitable.

It is even more problematic with the addition of special effects (e.g., graphics addons, beautification, cartoonisation) introducing non-natural features in natural content or pre-processing algorithms (contrast enhancement, sharpening, and noise-reduction) correcting most current artifacts in UGC videos.

Most existing full-reference quality metrics assume that the original content uploaded by users is the best quality and they have been tailored and tuned to quantify the perceptual quality of natural content. However, they do not perform well on UGC due to the lack of pristine reference. In reaction, several no-reference metrics were developed for UGC [1]. Also, aesthetic metrics are potentially suitable for "beauty" enhancements.

To the best of our knowledge, current large-scale UGC datasets, such as Youtube UGC [2], KoN-ViD-1k [3], The ICME 2021 Grand challenge UGC dataset [4] and LIVE-VQC [5], do not embed filtered or enhanced content. New subjective quality experiments dedicated to UGC with special effects need to be conducted. With the help of this new ground truth, the most promising quality metrics could be fine-tuned or new quality metrics could be developed.

Accordingly, there are three main contributions of this work: (1) the designed content selection that precedes a pilot subjective quality experiment, (2) a benchmark on existing quality metrics for UGC with special effects, and (3) the fusion of existing quality metrics tailored to UGC videos with special effects. The development will follow the same structure and will come to an end with a conclusion on findings.

2. SUBJECTIVE EXPERIMENT

We have access to 93 UGC 10-second video sequences coming from the social media Douyin (Chinese version of TikTok). Media were processed with in-house algorithms, including Artifact Removal (AR), Noise Reduction (MCTD) and Sharpening with four resolutions and 16 Constant Rate Factor (CRF) conditions for h264 and h265 with and without Region Of Interest (ROI) encoding. Table 1 summarizes the information relative to the content set.

Content Selection: We conducted a multi-codec, multi-quality-metric, multi-scale, multi-pre-processing content selection through rate-distortion (B-D) clustering [6]. We first selected content through an 8-clustering taking into account B-D rate behaviors of 720p content encoded with h265. Two sequences were extracted from each category, with priority given to 5 aesthetically filtered videos. On the 16 se-

Table 1. Information summary about the collected content set.

	Provided content	Selected content
PVS	-	309 (3-10 per SRC)
CRF	16 : {16, 18, 20, 22, 24, 25, 26, 27, 28, 29, 30, 32, 34, 36, 38, 40}	3
SRC	93	16
Encoding resolutions	720p (720x1280), 540p (576x1024), 480p (480x854), 360p (360x640)	720p, vertical, except 9 PVS (3 SRC x 3 remaining Resolutions)
Pre-processing	6: {mctd-ar-sharp-low, mctd-ar-sharp-med, mctd-ar-sharp-high, mctd-sharp-low, ar-sharp-low, ar-mctd-sharp-low}	
Codecs	4: h264 and h265 with and without ROI encoding	
Metrics	29: SSIM, PSNR and VMAF variants, internal metric: VQScore, BRISQUE, NIQE, ILNIQE, VIIDEO, VSFA, VIDEVAL, RAPIQUE NIMA, NIMA$_mp$, MLSP	

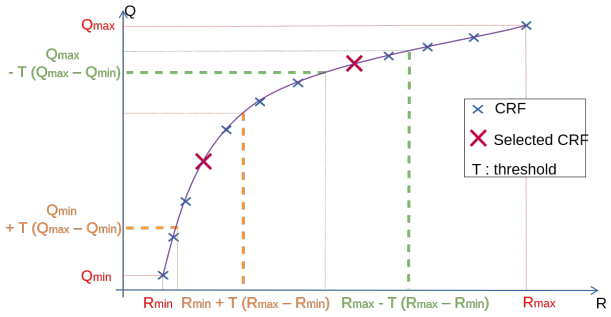

Fig. 1. Rate-quality optimized strategy to select CRF.

lected SRC content, we conducted a 17-clustering to select Hypothetical Reference Circuits (HRCs). Both clusterings consider VMAF-video bitrate encoding behaviors.

To select pre-processing HRCs, we ordered them by importance for the study, i.e., almost lossless content, ROI-based encoding, and no differentiation for remaining pre-processings. Indeed, it is necessary to include almost-pristine content (before delivery pipeline) first, then focus on ROI encoding, and finally, add Processed Video Source (PVS) with disruptive BD-rate behaviors. With respect to the prioritization, we keep only one PVS per SRC per cluster.

We envisioned several strategies for CRF selection: (1) constant selection (22, 26, and 32), (2) adaptive selection on the full scale, and (3) adaptive selection on the reduced scale 22-32. By adaptive selection, we considered a new approach to represent content behaviors towards encoding correctly. Indeed, taking the maxima or mean of the B-D curve based on one of quality or rate may not be fair. The mean CRF is the value found between the means bitrate- and quality-wise. Regarding the maxima CRFs, we define a threshold, expressed as the percentage of quality and bitrate ranges, defining the area the CRF falls into. We compute the bitrate and quality thresholds and select the CRF value that has bitrate and quality closest to the average of the defined range. This process is illustrated in Figure 1. Ultimately we proceeded with the full-range strategy (2) exhibiting CRFs between 24 and 32.

Subjective test design: We have conducted a crowdsourcing Absolute Category Rating (ACR) experiment on the Prolific platform. We defined 20 well-balanced playlists of 5 minutes, containing between 15 to 16 PVS. There are two balanced sets of population, one of which gathers Chinese speakers to study a population effect towards content presenting Chinese characters. Ultimately, we gathered 40 scores per PVS, i.e., 800 people participated in the test. In the following, when talking about the ground truth, we are referring to the collected ACR scores.

3. BENCHMARK OF QUALITY METRICS

We consider 29 off-the-shelf and most typical quality metrics in this benchmark, considering full-reference, blind metrics, and aesthetic models that may be specifically efficient on enhanced UGC videos.

Typical full-reference video quality metrics (9x2): Video Multimethod Assessment Fusion (VMAF) [7] efficiently predicts the quality of natural videos thanks to the fusion of fidelity, details, and motions features using a Support Vector Regression (SVR). Peak-to-Signal Noise Ration (PSNR) and Structural SIMilarity (SSIM) are state-of-the-art metrics based on signal noise and similarity. They present a trade-off between prediction accuracy and computational cost. Four versions of PSNR and SSIM are available, such as the global and three YUV channels representations (referred to as g-, y-, u- and v-metric). We need to tackle the lack of pristine original content to play the role of reference in full-reference metrics. We consider two references: the original non-pristine content, and the pre-processed and almost losslessly compressed (CRF1). It leads to two versions of these metrics results, for instance, u_PSNR and $crf1\_u\_PSNR$.

No-reference and blind metrics for Image/Video Quality Assessment (7): Instead of extracting distortion-specific, the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [8] quantifies visual losses utilizing the scene statistics. Similarly, Natural Image Quality Evaluator (NIQE) [9] models the spacial natural scene statistics without prior knowledge on content distortions, and its Integrated Local

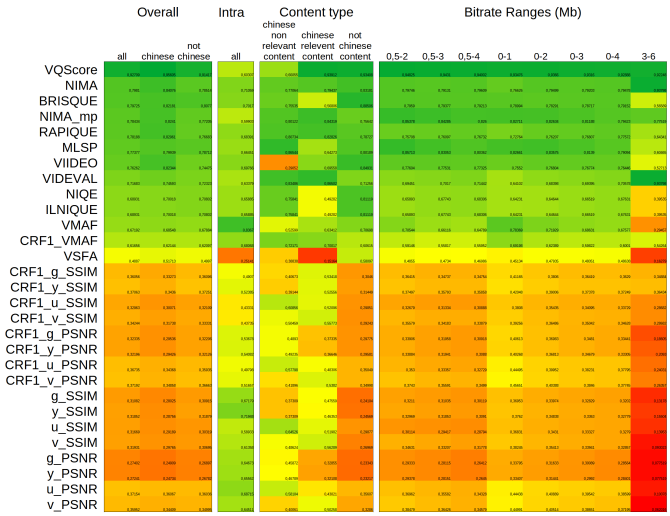

Fig. 2. Benchmark results of the C0 figure of merit on the 29 video quality metrics.

version (ILNIQE) [10] takes the prediction at a local scale through a spatial multivariate Gaussian model. The Video Intrinsic Integrity and Distortion Evaluation Oracle (VIIDEO) can transform disturbances between spatio-temporal sub-bands into predicted quality judgments in real-time. VSFA [11] models content- and memorability-dependant quality using a Convolutional Neurone Network (CNN) and Gated Recurent Unit (GRU). VIDEVAL is designed to better assess the quality of UGC by fusing different quality evaluators. Recently, Tu *et al.* proposed the Rapid and Accurate Video Quality Evaluator (RAPIQUE) [12] to evaluate the quality of UGC, which fuses the quality-aware scene statistics features with semantics-aware deep convolutional features.

Metrics for aesthetic assessment (3): The Neural IMage Assessment (NIMA) [13] is commonly considered as a baseline model. It is the first metric that predicts the aesthetic score via predicting the distribution of the ground truth data. A multi-patch pooled version (NIMA_mp) is also available. As global pooling is conducive to arbitrary high-resolution input, MLSP [14] is based on Multi-Level Spatially Pooled features. Even though some state-of-the-art models achieve appealing performance, none are designed for UGC videos with visual effects enhancements.

Additionally, we had access to the predictions of the VQScore (Video Quality Score) [15], an in-house proprietary metric of Bytedance designed and tailored to UGC.

Benchmark design: The benchmark relies on four typical linear figures of merit Pearson Linear Correlation (PLCC), Spearman Ranking Order Correlation (SROCC), Kendall Rank Correlation (KRCC), and Root Mean Square Error (RMSE), and three indicators from the Krasula's framework [16] in the pairwise paradigm. When evaluating the metric ability to discriminate and rank stimuli from a pair, it takes into account the uncertainty of subjective scores and is independent of the quality range of stimuli. It estimates the Area Under the Curve (AUC) regarding the discrimination (AUC DS) and rank (AUC BW) of paired stimuli, and the ranking correct classification rate, C0. We report here only results with C0, but the same conclusions exhibited from other criteria.

We investigated the behaviors of quality metrics towards different influence factors: the population (Chinese speaker or not), the content type (contains Chinese references or characters), the bitrate and quality ranges, the consideration of intra or inter pairs. Differentiating intra and inter pairs displays the specific ability of a metrics to discriminate stimuli from the same source or not, respectively. The intra-pair pipeline is a comparison fairest to full-reference metrics designed to compare same-source content.

Benchmark results: Figure 2 introduces the results per population, intra, content type, and bitrate ranges. The colors indicate the C0 scores obtained by metrics: green and red represent 1 (perfect classification) and 0, respectively. Metrics VQScore, NIMA, BRISQUE, NIMA_mp, RAPIQUE, MLSP, VIIDEO are the most performing metrics, in this order. VQScore is by far the best, showing it is already well-tailored to enhanced and filtered UGC. Note the mild accuracy of full-reference metrics overall, together with VSFA. As expected, full-reference metrics are far better on intra pair comparisons, where blind and aesthetic metrics may suffer from the lack of prior knowledge about the content. Several metrics are noticeably highly performing in high bitrates (VQScore, NIMA, VIDEVAL and to a lesser extent and NIMA_mp). Finally, there is relatively no effect on the population.

4. COMBINATION OF METRICS

To obtain more efficient and suitable video quality metrics for UGC with special effects, we regard each quality assessment model selected in the benchmark and fine-tune a machine learning model to fusion them in a better predictor. Due to the expensive cost of feature extraction, we want to find a low-dimensional and lightweight group of metrics combined in a highly performing new metric. Fine tuning such a metric on the new dataset makes it specific to enhanced and filtered UGC.

Fast convergence strategy: Fine-tuning machine learning techniques that fuses any combination of the 29 quality metrics is too demanding. Instead, we defined a strategy converging quickly to the most satisfactory combination of metrics. Backward Feature Elimination (BFE) is a practical greedy dimensional-reduction method that particularly fits our use case. From the initial set of metrics, it removes the least performing feature at every iteration. By iteration, we mean that at step k, we compute the combination of metrics following a leave-one-out strategy. The discarded metric/feature is the one missing from the most performing fu-

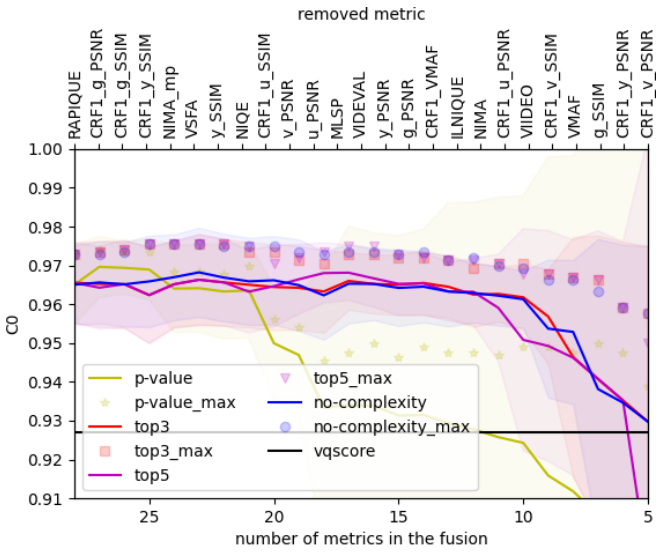

Fig. 3. BFE convergence from the fusion of 29 metrics to 5. It shows the average and standard deviation performance of all combinations, and the most performing one, easily compared to the performance of VQScore.The upper axis indicates the metric removed at each iteration for no-complexity strategy.

sion of metrics. The stopping rule is to keep only 5 features. We will verify empirically that this threshold is meaningful.

Ultimately, we will need to choose a combination of metrics among all the selected ones during the BFE. To select the final fusion of metrics we calculate the ratio of performance and complexity and keep the fusion of metrics with the highest ratio.

Performance and complexity criteria: We look for a fusion of metrics that reaches the fine balance between accuracy and complexity to match the constraints of broadcasters (performance evaluations must not be more complex than the encoding). First, we select the C0 exhibited by the Krasula framework as a performance indicator. It focuses on the correct classification rate of metrics when discriminating pair comparisons.

Then, we define four strategies to achieve the fine balance between performance and complexity: (1) ***no-complexity***: the performance indicator C0 evaluates alone the fusions of metrics. (2) ***p-value***: this strategy rejects the model not included in the most complex fusion among the range of combinations that have similar performance (i.e., not proven significantly different by a t-test). (3) ***Top 3***: finds the three best performance combinations and deletes the metric that has the highest complexity. (4) ***Top 5***: similar to the above, but considers five top performing features.

The complexities of quality models are set based on their computation requirements (i.e., PSNR, SSIM:1; VMAF:2; RAPIQUE, VQScore:3; BRISQUE:4; NIQE:5, VIDEVAL:6; ILNIQUE:7; VIIDEO:8; MLSP, NIMA, NIMA_mp, and VSFA:9).

Table 2. Performance of the selected combination of metrics, when compared to VMAF and VQScore. Best results are highlighted.

	Metrics fusion	VQScore	VMAF
PLCC	**0.9339**	0.8632	0.3496
SRCC	**0.9297**	0.8597	0.3734
KRCC	**0.7700**	0.6672	0.2544
C0	**0.9656**	0.9271	0.6719

Fusion of metrics Numerous solutions can fuse metrics such as decision trees, random forests, AdaBoost, SVR or a simple CNN regressor. We performed a grid search to select the algorithm (among SVM (classification), SVR (regression), Random forest, and Adaboost) and its hyper parameters (e.g. kernel, value of kernel) on the first BFE iteration. In the following, we train an SVR for each metrics fusion, the best predictor based on the gird search. We adopt a 5-fold cross-validation process with an allocation of 80% and 20% to training and test sets, respectively, with a grid search to tune the classifier.

Results We consider the 29 quality metrics involved in the benchmark and use the ACR scores as ground truth. Figure 3 shows the convergence of the BFE for all four strategies. Removed metrics for the no-complexity strategy are given on top of the figure. We can see that the p-value strategy impacts too drastically the efficiency of the fusions. The remaining strategies are equivalent, with the no-complexity strategy that presents a less drastic drop around 5-metrics combinations. This strategy is thus selected in the following. The final combination of metrics is an SVR RBF kernel with gamma at 0.001 with a regularization parameter set to 10 with seven dimensions, i.e., BRISQUE, VQScore, v_SSIM, u_SSIM, g_SSIM, CRF1_y_PSNR, and CRF1_v_PSNR. Its performance is depicted and compared to VQScore and VMAF in Table 2. Mixing two of the most performing low-complexity metrics (VQScore and BRISQUE) with full-reference low-complexity metrics (PSNR, SSIM) benefit from intra and inter qualitative predictions.

5. CONCLUSION

The urge to specialize video quality metrics is decisive for specific use-cases such as UGC enhanced and filtered videos. We have designed an intelligent content selection strategy, collected ACR scores in crowdsourcing, ran a benchmark on current full-reference, blind and aesthetic metrics. We finally introduced a methodology to tailor a performance-complexity fusion of metrics. The exhibited tailored metric is in line with the benchmark findings: VQScore is of high quality, for all bitrate ranges and, with BRISQUE, benefits from full-reference low-complexity metrics to improve the intra-pairs predictions.

6. REFERENCES

[1] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[2] Yilin Wang, Sasi Inguva, and Balu Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[3] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, "The konstanz natural video database," 2017.

[4] Haiqiang Wang, Gary Li, Shan Liu, and C.-C. Jay Kuo, "Challenge on quality assessment of compressed ugc videos," 2021.

[5] Zeina Sinno and Alan C. Bovik, "Large scale subjective video quality study," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 276–280.

[6] Suiyi Ling, Yoann Baveye, Patrick Le Callet, Jim Skinner, and Ioannis Katsavounidis, "Towards perceptually-optimized compression of user generated content (ugc): Prediction of ugc rate-distortion category," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[7] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock, "Vmaf: The journey continues," *Netflix Technology Blog*, vol. 25, 2018.

[8] Peng Ye and David Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129–3138, 2012.

[9] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[10] Lin Zhang, Lei Zhang, and Alan C Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[11] Jari Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[12] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik, "Rapique: Rapid and accurate video quality prediction of user generated content," *arXiv preprint arXiv:2101.10955*, 2021.

[13] Hossein Talebi and Peyman Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[14] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 9375–9383.

[15] Yang Li, Longtao Feng, Jingwen Xu, Tao Zhang, Yiting Liao, and Junlin Li, "Full-reference and no-reference quality assessment for compressed user-generated content videos," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[16] Lukáš Krasula, Karel Fliegel, Patrick Le Callet, and Miloš Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.