When is the cleaning of subjective data relevant to train UGC Video Quality Metrics?

Anne-Flore Perrin, Charles Dormeval, Yiling Wang, Neil Birkbeck, Balu Adsumilli, Patrick Le Callet

WHEN IS THE CLEANING OF SUBJECTIVE DATA RELEVANT TO TRAIN UGC VIDEO QUALITY METRICS?

Anne-Flore Perrin[§] *Charles Dormeval*[§] *Yilin Wang*[†] *Neil Birkbeck*[†] *Balu Adsumilli*[†]
Patrick Le Callet[§]

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

ABSTRACT

Outlier analysis and spammer detection recently gained momentum in order to reduce uncertainty of subjective ratings in image & video quality assessment tasks. The large proportion of unreliable ratings from online crowdsourcing experiments and the need for qualitative and quantitative large-scale studies in the deep-learning ecosystem played a role in this event. We study the effect that data cleaning has on trainable models predicting the visual quality for videos, and present results demonstrating when cleaning is necessary to reach higher efficiency. To this end, we present and analyze a benchmark on clean and noisy User Generated Content (UGC) large-scale datasets on which we re-trained models, followed by an empirical exploration of the constraint of data removal. Our results show that a dataset presenting between 7 and 30% of outliers benefits from cleaning before training.

Index Terms— Outlier, Cleaning, Video Quality Metrics, User Generated Content, Training metrics

1. INTRODUCTION

This last decade, Video Quality Metrics (VQMs) based on deep neural networks [1, 2, 3] reached higher efficiencies than previous hand-crafted features solutions. However, these models often require large datasets for training and validation. This requirement led to the curation of large-scale representative image and video datasets with corresponding human subjective ratings, which are well beyond what was possible to acquire using in-lab study designs. Crowdsourced data collection was the solution to scale-up the datasets [4, 5], but the approach came with additional challenges in terms of outlier and spammer removal. Traditional subjective screening protocols (e.g., ITU BT500 [6]) are not applicable, as not all subjects see all video stimuli in crowdsourced experiments, implying fewer overlapping stimuli between raters, limiting inter-rater consistency checks.

Advanced data-cleaning methods have been developed (e.g., [7, 8]), but their efficacy has not been evaluated on such large scale crowdsourced datasets. Furthermore, despite outlier removal and subject screening being standard practice, the impact of outliers on Mean Opinion Score (MOS) and the resulting effect on learned video quality metrics is largely unstudied. For instance, it is unclear how beneficial screening and removing unreliable data will be for learning-based models, and at what proportion of outliers are the models affected.

In this paper, we will first explore and analyze several state-of-the-art probabilistic data-cleaning methods [7, 8] on a large-scale subjective video quality datasets: the YouTube UGC Dataset [4] H.264. We evaluate these methods with respect to their ability to improve confidence in our estimates, and test repeated applications of these methods to understand their convergence (Section 2.2). We then perform extensive analysis of several state-of-the-art learned blind VQMs and evaluate their performance with several criteria (Section 3), including typical correlation figures of merit and pairwise-paradigm evaluations that take into account uncertainty of the subjective scores [9]. When varying the proportion of outliers in the training dataset, we confirm that the more outliers, the stronger the need is to screen out unreliable behaviors (Section 4). One of the main contributions is the analysis methodology used to evaluate the sensitivity of VQM training to outliers in the training set.

2. APPLYING OUTLIER AND SPAMMER DETECTION

Our analysis focuses on UGC video, which has gained exponential growth with the daily use of social media. Such content, created by casual users, represents a large share of the internet traffic. The quality of UGC videos may be affected by multiple distortions (e.g., compression, blur, noise) and aesthetic issues–causing differing opinions on overall quality. For this reason, UGC content is challenging to get accurate ratings, and, together with online tests, it serves as a practical testbed to evaluate the impact of outliers on VQM efficacy.

We focus on YouTube UGC H.264 dataset (YT-UGC MOS) [4], collected on 9544 online observers, each assigned

Table 1. Results of outlier detections: number of detected outliers ($\#o$) over the $N = 9544$ raters, D_{MOS}, and D_{CI} when compared to the previous iteration. Same information for the combination of metrics. Underlined results indicate a positive gain in confidence (reduction of CIs).

		iteration					models		
algorithm		1	2	3	4	5	combination		(1 + 2) overall
	#o	<u>872</u>	+34	+34	+35	+36		#o	(872 + 191) <u>1063</u>
MLE CO	D_{MOS}	0.0603	0.0037	0.0041	0.0042	0.0039	MLE CO + GPM	D_{MOS}	0.072
	D_{CI}	14.44	0.433	0.403	0.431	0.421		D_{CI}	15.42
	#o	<u>664</u>	<u>+11</u>	<u>+1</u>				#o	(664 + 387) <u>1051</u>
GPM	D_{MOS}	0.0654	0.0031	0.001			GPM + MLE CO	D_{MOS}	0.066
	D_{CI}	14.51	0.195	0.025				D_{CI}	12.51

with a playlist of 30 video stimuli over 1385 sequences overall. Participants rated continuously on an Absolute Category Rating (ACR) scale from 1 to 5. Some outlier detection algorithms only support discrete ACR scores; therefore, we performed quantization by rounding. This assumption is valid in that if only a discrete ACR scale was used, the rater would have picked the nearest integer score.

For ecological validity, it is good practise to perform the analysis on several datasets. However, it is very difficult to find datasets suitable to our needs: some do not share raw scores but only report MOS and CIs (e.g. LIVE-VQC [10]) ; some already removed screened outliers (e.g; KoNViD-1K [5]); some are not large enough for metric fine-tuning (e.g. LIVE Netflix VQoE [11]); some are unreliable or do not provide enough information about the dataset production (e.g. ICME grand challenge QAC UGC [12]).

As our analysis in subsequent sections depends on the outlier detection, and since there is no ground truth for outlier detection, we pay special attention to review (§2.1) and to evaluate the most relevant options (§2.2) below.

2.1. Outlier Detection methods

For in-lab studies, the International Telecomunication Union (ITU) standardised a method based on box-plot outlier detection [6]. Numerous other techniques tackle the issue such as density-based algorithms, with Majority Voting (MV) [13] - naively assuming the most answered label is the ground truth - or the Probability Modeling (PM) [13] - using Expectation-Minimization (EM) to derive workers' quality. Recently, elegant solutions rely on Maximum Likelihood Estimation (MLE) like the MLE CO [14] inferring a subject's bias and inconsistency. Other models use graphical models [8] and EM to infer reliability and answers regularity (GPM) [7] or annotator expertise, content ambiguity and most probable label (GLAD) [15]. These techniques apply on non-binary linear-scale scores, i.e. for ACR or Double Stimulus Impairment Scale (DSIS) scores. In case of binary answers, such as the expression of preference during a Pair Comparison (PC) test, recent works [16, 17] favored the use of dissimilarity metrics (Rogers-Tanimoto (RT) dissimilarity [18], Cohen's Kappa Coefficient [19]) to estimate inter-observer agreement.

All outlier techniques are not applicable on large-scale datasets. For instance, crowdsourcing collections rarely present raters' playlists with sufficient intersection to conduct dissimilarity analysis. Thus, PM and binary outlier detections have been dropped from the study. GLAD is computationally costly. As it is conceptually close to GPM, we do not include it in the study. MV suffers from the simplicity of its model and its prior assumption that all raters are reliable, and thus penalises too much good raters. Finally, Youtube UGC already rejected outliers detected with algorithms form the ITU. For these reasons, we consider the GPM and MLE CO techniques in the following analysis.

2.2. Outlier Detection Evaluation

We defined two figures of merit to verify the outlier removal efficiency. We want to verify that MOS are relatively stable and that 95% Confidence Intervals (CIs) have been reduced. To study the amount of variation introduced by the detection, we consider the Root Mean Square Error (RMSE) between MOS before and after the screening, and the sum of the absolute difference between "clean" (without scores from outliers) and "noisy" CIs:

$$D_{MOS} = RMSE(MOS^{noisy}, MOS^{clean})$$
$$D_{CI} = \sum_{i=0}^{N_{stim}} |CI_i^{noisy} - CI_i^{clean}|$$

with N_{stim} the number of stimuli.

The MLE CO algorithm rejects raters with a bias and inconsistency over a specific threshold. We empirically set the threshold to 1, which corresponds to 5% of raters showing disruptive behaviors. Interquartile range (IQR) [16] or other techniques could have been used interchangeably. Also, note that the GPM method assumes that most raters are reliable.

We first observe that both methods do not necessarily detect the same raters as outliers. For instance, there is an intersection of 477 outliers between GPM and MLE CO. We thus decided to explore repeating or cascading the outlier detections, as such a process may end up screening outliers more efficiently. Table 1 introduces the number of outliers and the figures of merit results for repeating and cascading models. Repeating GPM always reduces CIs, i.e., increases our confidence in scores, and stops finding outliers after three iterations. Note that the first iteration removed most of the ob-

Table 2. Efficiencies of 5-fold 60/20/20 trained metrics on clean and noisy data. Numbers in brackets are the std over the folds. We highlighted the best results per figure of merit, per VQM. Underlined figures means a significant difference has been found by a statistical test.

VQMs	SRCC↑	KRCC↑	PLCC↑	RMSE↓	AUC D-S↑	AUC B-W↑	C0↑
RAPIQUE clean	**0.769** (0.014)	**0.574** (0.01)	**0.775** (0.01)	**0.387** (0.014)	**0.680** (0.002)	**0.907** (0.0005)	**0.826**
RAPIQUE noisy	0.765 (0.013)	0.572 (0.011)	0.771 (0.011)	0.392 (0.011)	0.679 (0.002)	0.905 (0.0005)	0.825
VIDEVAL clean	**0.734** (0.041)	**0.537** (0.037)	0.721 (0.046)	0.427 (0.03)	**0.664** (0.002)	**0.8879** (0.0006)	**0.807**
VIDEVAL noisy	0.733 (0.041)	0.536 (0.036)	**0.722** (0.044)	**0.424** (0.022)	0.663 (0.002)	0.8874 (0.0006)	0.806
TLVQM clean	**0.596** (0.026)	**0.419** (0.023)	0.593 (0.034)	**0.481** (0.013)	0.592 (0.002)	0.8207 (0.0007)	**0.7422**
TLVQM noisy	0.595 (0.028)	0.417 (0.024)	**0.594** (0.037)	0.482 (0.014)	**0.593** (0.002)	**0.8208** (0.0007)	0.7419
VSFA clean	**0.752** (0.02)	**0.553** (0.017)	**0.752** (0.021)	**0.409** (0.025)	**0.651** (0.002)	**0.897** (0.0006)	**0.812**
VSFA noisy	0.737 (0.028)	0.538 (0.024)	0.737 (0.028)	0.427 (0.042)	0.641 (0.002)	0.887 (0.0006)	0.802

servers presenting disruptive behaviors. Per design, MLE CO always finds outliers, but it does not increase the reliability of mean scores after a first pass. We thus concluded that applying the algorithm once is sufficient.

Accordingly, we tried the combinations of methods. Reasonably, the combination MLE CO + GPM gains more in reliability (higher D_{CI}) with similar conservation of MOS. Note that this result is in line with the GPM underlying assumption.

Overall, with the MLE CO + GPM pipeline, we detected 1063 spammers over the 9544 annotators, i.e., 11% of raters are identified as unreliable. It confirms the high quality of the YT-UGC MOS dataset, especially regarding the crowdsourcing collection. In the following, we refer to MLE CO + GPM when talking about the outlier detection.

3. BENCHMARK OF BLIND VIDEO QUALITY METRICS

We then conduct a benchmark study to understand the influence of outlier removal on the training of VQMs. We first clean the data and provide clean and noisy datasets to train new models through a 5-fold cross-validation (using 60/20/20% splitting into training, validation, and test sets). We train each quality model on clean and noisy data, and evaluate the predictions of metrics on the clean test set. We call a metric *clean* (resp. *noisy*) when it has been trained on a dataset without (resp. with) the rejected scores of unreliable raters.

We consider the four blind VQMs mostly used in the UGC ecosystem [1], namely RAPIQUE [2], VIDEVAL[1], TLVQM [20] and VSFA [3]. Briefly, VSFA implements a Convolutional Neural Network (CNN) and Gated Recurent Unit (GRU) scheme modeling content- and memorability-dependant quality. VIDEVAL selects the best features from state-of-the-art metrics through random forest and fusion them with a Support Vector Regression (SVR). TLVQM considers local temporal quality (chunks of 1 second). TLVQM identifies the most representative frame of a chunk through 22 low complexity features (LCF) and computes 30 high complexity features (HCF). An SVR is then trained on averages and standard deviations (std) of frames LCF and chunks HCF. Finally, RAPIQUE combines low-complexity spatio-temporal features along with a semantic-based CNN.

To conduct a fair comparison between video quality metrics, and their clean and noisy versions, we used several indicators: the typical Pearson Linear Correlation (PLCC), Spearman Ranking-order Correlation (SRCC), Kendall Rank Correlation (KRCC), and RMSE, and measure statistical significance between the distributions by a t-test. To look further into the influence of unreliable raters into training sets on metrics, we include Krasula's framework [9] operating in the pairwise paradigm. It expresses metrics' discriminating power (whether stimuli from a pair show a significant difference) and ranking power (assessing which stimuli is better). It takes into account the uncertainty of subjective scores and is independent of the quality range of stimuli. From Krasula's framework [9], we extract the Area Under the Curve (AUC) D-S between the Different and Similar distributions (discriminating power), the AUC B-W between the Better and Worse distributions (ranking power), and its percentage of correct classification (C0). In order to assess statistical difference between metrics, the Fischer exact test is used for C0, and the Hanley-McNeil power law for AUC.

The results of the benchmark are in Table 2. For all linear figures of merit but PLCC, screening and removing outliers before training a metric is beneficial. The RMSE for VIDEVAL also shows significantly better noisy version. It hints that this metric is particularly robust to outliers. No significant differences were found in the pairwise measures (e.g., AUC D-S, AUC B-W, C0).

From these results, although outlier detection does not appear to be strictly beneficial, it also does not degrade the efficiency of the trained metrics. Therefore, we still recommend to always apply outlier detection.

We expected significantly lower performance of the noisy metrics, YT-UGC MOS is of high quality with only 11% of detected outliers. We suspect having hit the limit of the outlier screening gain due to the variations in correlations and RMSE and the lack of statistical significance in pair-wise figures of merit.

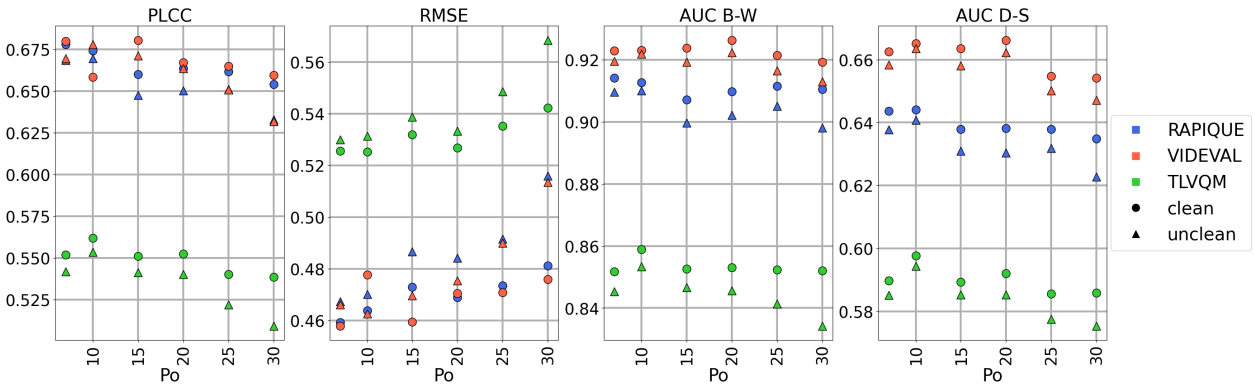

Fig. 1. Evolution of clean and noisy metrics learnt on datasets of various outlier percentage (p_o) and $N = 15$ in function of figures of merit scores.

4. IMPACT OF OUTLIER RATIO ON TRAINABLE METRICS

We are now interested in understanding the relationship between the outlier ratio and metric efficiency. The YT-UGC MOS dataset is currently the only dataset that enables such a study, to the best of our knowledge. We assume that the outliers detected by MLE CO + GPM are the only outliers. Let us define the number of votes per stimulus N, the number of votes given by outliers per stimulus N_o, giving as percentage of outliers $p_o = \frac{N_o}{N} \times 100$.

We examine p_o over the range [7, 30] by dichotomy and N with values within $\{15, 20, 30, 50, 70, 90, 100\}$. We proceed with a two-step 5-fold cross-validation pipeline: (1) to generate a sub-dataset having the desired properties, i.e., creating 5 batches with N_o random scores of outliers per stimulus, each concatenated with a different random set of $N - N_o$ reliable scores, and (2) to train metrics on each of the created sub-datasets. Note that by taking randomly reliable outliers, we lose the consistency in a subject's expectations, which may slightly bias the process. We quantify the low variations between sub-datasets and the ground truth with the RMSE between full- and sub- datasets' MOS, and with p_o = 30. We happily obtain differences under 0.08.

There may not be up to N_o unreliable scores for some PVS, for which we must simulate outliers' scores. We thus draw ratings from the scores distribution of a random outlier. We define as disruptive score $dis(s) = \sum_i |MOS(s) - sim_i(s)|/M$, with $sim_i(s)$ the i^{th} simulated outlier score of stimulus s. The original dataset with ground truth outliers reaches a disruptive score of 1.16. For all tested configurations $\frac{N}{p_o}$, the average disruptive score for simulated outliers is similar (1.34), implying the simulation was effective.

In Table 3 and Figure 1, we illustrate the evolution of metrics efficiency depending on p_o and N for all figures of merit. As expected, we observe that outlier removal is always beneficial, especially at a high percentage of outliers, particularly for RMSE. We also note that exploiting subjective scores of a high number of PVS may overcome the noise introduced by outliers when the dataset contains just a few of them (e.g. 7% p_o and $N > 30$, with $0.05 <$p-value< 0.1). RAPIQUE seems to be the least robust to outliers as the delta between its clean and non-clean versions is larger than the other metrics for all outlier percentages, contrarily to VIDEVAL. This effect is mild but enables metrics comparisons. This delta could be used as a proxy for metric robustness to outliers.

Table 3. Delta RMSE between RAPIQUE clean and not clean metric. Variations over number of raters per PVS N and percentage of unreliable scores p_o ; underlined deltas indicate a t-test p-value < 0.05, other values are < 0.1

p_o \ N	15	30	50	70	90	110
7	**0.0079**	**0.004**	0.0031	0.0023	0.0024	0.0015
15	**0.0136**	**0.0112**	**0.0086**	**0.0079**	**0.0073**	**0.0075**
21	**0.0153**	**0.0184**	**0.0152**	**0.0157**	**0.0133**	**0.0154**
30	**0.0348**	**0.0265**	**0.0277**	**0.0268**	**0.0299**	**0.0311**

5. CONCLUSION

In this work, we define how to combine several outlier detection models to detect most outliers. Cascading the MLE CO and the GPM best reduces subjective uncertainty while keeping MOS values stable.

In a benchmark of VQMs learned on clean and noisy datasets, we verified that applying outlier detection is at worst transparent, and at most beneficial. We also confirmed that the benefits of screening increases with higher outlier rates. The designed study allows identifying which metrics are robust to outliers, which is a highly interesting outcome.

In the future, we aim to extend and validate the results on other suitable datasets. Also, it would be good to extend the analysis to binary-choices methodologies (e.g., RT dissimilarity [18]).

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] 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.

[3] 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.

[4] 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.

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

[6] ITU Recommendation BT.500-14, "Methodologies for the Subjective Assessment of the Quality of Television Images," 2019.

[7] Jing Li, Suiyi Ling, Junle Wang, and Patrick Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3339–3347.

[8] Yingtong Dou, "A review of recent advance in online spam detection," -, 2019.

[9] 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.

[10] 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.

[11] CG Bampis, Z Li, AK Moorthy, I Katsavounidis, A Aaron, and AC Bovik, "Live netflix video quality of experience database," *Online: http://live. ece. utexas. edu/research/LIVE NFLXStudy/index. html*, 2016.

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

[13] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng, "Truth inference in crowdsourcing: Is the problem solved?," *Proceedings of the VLDB Endowment*, vol. 10, no. 5, pp. 541–552, 2017.

[14] Zhi Li, Christos G Bampis, Lucjan Janowski, and Ioannis Katsavounidis, "A simple model for subject behavior in subjective experiments," *Electronic Imaging*, vol. 2020, no. 11, pp. 131–1, 2020.

[15] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," *Advances in neural information processing systems*, vol. 22, pp. 2035–2043, 2009.

[16] Ali Ak, Mona Abid, Matthieu Perreira Da Silva, and Patrick Le Callet, "On spammer detection in crowdsourcing pairwise comparison tasks: Case study on two multimedia qoe assessment scenarios," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[17] Mary L McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[18] David J Rogers and Taffee T Tanimoto, "A computer program for classifying plants," *Science*, vol. 132, no. 3434, pp. 1115–1118, 1960.

[19] Jacob Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[20] 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.