https://doi.org/10.2352/EI.2024.36.11.HVEI-216


# Study on the relationship between quality and acceptability annoyance (AccAnn) of UGC videos

***Pierre David*[1,2], *Pierre Lebreton*[1,2], *Neil Birkbeck*[3], *Yilin Wang*[3], *Balu Adsumilli*[3], *Patrick Le Callet*[1,4]**
[1] ***Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, Nantes, France,*** [2] ***CAPACITÉS SAS,*** [3] ***Google Inc.,*** [4] ***Institut universitaire de France (IUF)***

## Abstract

*We study the relationship between acceptability/annoyance rating and traditional MOS quality ratings of UGC videos. Acceptability/annoyance is a key concept for evaluating services, as it classify whether the delivered service quality falls into acceptable, annoying but acceptable, or not acceptable. This relates to the willingness of users to use those services. While audiovisual quality estimation models have a long research history, the translation of these quality scores to acceptability and willingness to use the services has only been weakly studied. In this work, a new dataset was then created to evaluate both quality and the acceptability/annoyance of videos. Different state-of-the-art quality prediction models videos were evaluated at predicting quality of UGC videos. Furthermore, performance at predicting acceptability/annoyance of videos was also tested.*

## Introduction

User Generated Content (UGC) video streaming is one of the major application on the Internet. Indeed, video streaming represents a major part of the Internet traffic, with YouTube and Netflix combined represented already more than a fith of the internet traffic in 2021 [1].

Considering the volume of videos to stream, there is a need for perceptually-based encoding to allow decreasing the amount of data to be sent while preserving the quality of the videos. Netflix has described the use of quality metrics such as VMAF[2] for choosing bitrate ladders. This allows to consider the coding complexity of the contents [3, 4], and enable decreasing the required amount of data that needs to be sent. Therefore, there is a strong need to evaluate the quality of experience of users when watching UGC videos.

### *Quality evaluation of UGC videos*

When UGC videos are considered, new challenges arise as the videos distributed by the platform were created and uploaded by casual users. Due to this user-generated nature, the quality of the videos cannot be guaranteed. Furthermore, quality evaluation metrics such as VMAF [2] that was used to evaluate the different encode trials requires a pristine quality reference to estimate the quality of a degrated video. Therefore, in the context of UGC videos, VMAF cannot be reliably used to estimate the quality of transcoded videos.

Due to this need for reference agnostic video quality metrics, a large amount of effort has been put into the development of no-reference video quality metrics that do not require a pristine quality reference. Tu [1] describes a benchmark of different no-reference video quality assessment method. The considered metrics are based either on Natural Scene Statistics (NSS) features [5], dictionaries of distortions [6], multi-level handcrafted features [7], or with deep convolutional neural networks (CNN) features[8, 9]. Additionally, Tu [1] proposed a new feature integration method called VIDEVAL that aggregates selected features from previous NSS-based models.

Going into fully data-driven models, different video quality assessment methods have been proposed. Zhang [10] showed that deep features from CNN models, such as VGG16 that was trained on image classification task, can efficiently be used for quality evaluation purpose. Li [11] described a model that combines a "content-aware feature extraction" module that is based on a pre-trained CNN model and a "temporal-memory" based on a Gated Recurrent Unit (GRU) network for modeling temporal aspects. Xu[12] proposed a model that performs feature extractions using a Graph Convolution Network (GCN), evaluated the relative importance of features using an attention module, and temporally integrated features using a Bi-LSTM. Then, considering the recent advancement from Vision Transformers (ViT) [13], new attention-based quality evaluation models were proposed [14, 15]. These leverage ViT to model attention and weight the contribution of different spatial regions.

Going beyond quality evaluation Wang [16] describe the UVQ model that aims at better understanding quality predictions. The authors describe a "Comprehensive Interpretation Network for Video Quality" (CoINVQ) that is composed of thee sub-graphs that output three different key information: the type of content, the distortions induced during production of the content, and the compression level from coding. The output of the different sub-graph would then enable to better understand the construct of video quality ratings. Finally, a overall quality score is also computed using a temporal integration module.

### *Acceptance evaluation*

Although quality prediction has been intensively studied, from an operator point of view it is also important to know how to interpret the quality scale and identify what level of quality is acceptable for the user. Several studies have then addressed the question of acceptability. First, Jumisko-Pyykko[17] defines acceptability as a binary measure. The goal is to locate the threshold of minimum acceptable quality that fulfills user quality expectations and his needs for a certain application or system. The evaluation process of acceptability was defined in the ITU-T Recommendation P.10/G.100 stating that this scale should be evaluated on the basis of a Yes/No answer so to put a low

cognitive burden on the participants [18]. However, evaluating acceptability is not simple. Indeed, previous works have considered that what may be considered as an acceptable quality may not be enjoyable. Therefore, studies were conducted with different scales so to address different levels of acceptability. This was done by asking participants about not only "acceptability" but also "pleasing acceptability" by Song [19], or whether the quality level was "annoying" by Li [11]. The latter was done using an acceptability/annoyance (AccAnn) scale and enabled the steps leading from annoyance to unacceptability to be better understood. Finally, it should also be mentioned that works have aimed at measuring acceptability using the relationship between quality and and whether users wish to quit watching videos [20, 21].

Aiming at better understanding the relashionship between scales, Jumisko-Pyykko [17] studied the relationship between "pleasing acceptability" and "acceptable quality". She found that pleasing acceptability would be reached at a higher quality level than acceptable quality, and a three order polynomial mapping could be found between the two. Several works have then aimed at mapping the quality-related parameters such as QP, bitrate, PSNR and Structural Similarity Index (SSIM) to acceptability using sigmoidal functions [17, 22, 23, 24]. Koning [25] proposed to map NTIA VQM quality scores [26] with acceptability ratings using a three order polynomial function. Pessemier [27] described a model based on a decision tree with network, watching behavior, and video quality -features. Finally, Li [11] proposed a classification algorithm to map predictions from quality levels into AccAnn levels, and Ak [28] showed that depending on the context of evaluation the relationship the results on the relationship between quality and acceptability could vary.

### Contributions

As discussed above, significant work has been done on the evaluation of UGC videos' quality, and on the relationship between quality and acceptability. However, the acceptability of UGC videos has only been weakly addressed. Key research questions this work investigates are the following:

- How different UGC datasets relates with each others, and what are the consequences on video quality prediction models.
- How well the UGC quality prediction models can predict acceptability of UGC videos.

In the following, Section  describes a large scale crowdsourced experiment that evaluates both quality and acceptability of UGC videos. Section  provides the analysis on the two research question. Finally, Section  concludes this paper.

## Experiments

In this study, the relationship between quality and acceptability of UGC videos is addressed. To this aim, various content were presented at various levels of quality and were evaluated both in terms of quality and acceptability. This section describes the selection process of source content, the definition of coding conditions, and the subjective evaluation methodology.

### Content selection

In the literature, several UGC video datasets have been established (Table 1). These datasets contain videos with various quality due to being available at different resolutions, frame rate, and/or were captured using cameras of various quality. These videos are provided with subjective scores that can be based on ratings from observers. The number of ratings can be as low as 5 observers in the case of KoNViD-150K B [29], or as high as beyond 150 participants in the case of the YouTube-UGC dataset [30]. Figure 1 shows histograms of quality scores of the dataset that were available for download. The subjective score depicted in these bar graph were collected by the respective authors of these datasets. From this Figure, it can be seen that each dataset shows different quality ratings distribution. The YouTube-UGC dataset is tailored towards high quality videos, the KoNViD-1K and Live Qualcomm datasets show quality values centered around a MOS of 3. The KoNViD-150K is available in two version "A" and "B". In the case of KoNViD-150K.A 152 265 videos were evaluated by 5 observers, while in the case of KoNViD-150K.B 1577 videos were evaluated by at least 89 observers. Subjective score distribution shows that KoNViD-150K.B mostly shows average quality videos, while KoNViD-150K.A contains higher quality ratings. Finally, the YouKu-V1K and Netflix public datasets show rather uniform distribution of quality values.

However, an important aspect to note, is that all datasets are not necessarily aligned, and a quality rating "4" in the YouTube-UGC dataset does not necessarily correspond to a quality rating "4" in the Youku-V1K dataset. Indeed, as the context of evaluation, the type of quality impairments, and the range of quality values seen by participants is different in each experiment, and the ratings are then given in different referentials.

In order to study the relationship between quality and acceptability across a large variety of conditions, videos from these open datasets were selected. Doing so allows increasing the diversity of our test set in terms of contents, types of impairment, camera systems, etc., as each dataset has varying diversity in these dimensions. Secondly, this also allow learning the relationship between the quality scores from the different open datasets.

As the relationship between quality and acceptability is aimed, it is also proposed to include coding conditions by transcoding the videos taken from the open datasets into lower quality levels. The videos directly taken from the open dataset will then be referred as "ingested videos" as if these videos were directly uploaded by users onto a video streaming platform. These "ingested videos" are then transcoded to lower quality so to mimic the different step of a bitrate ladder.

Figure 2 depicts the quality of the "ingested videos" that were selected from 5 of the open datasets. The selection process was done so to cover the range of quality values from each dataset. Secondly, diversity in terms of type of content (gaming, travel, talk shows, etc.) was also aimed. As in this experiment, ingested videos are transcoded at various quality, an over-representation of high ingest quality was done so to allow the test to have balanced quality condition once coding conditions are added. Finally, the duration of the videos were reduced to five seconds. At the end of this process, 326 "ingest videos" were selected.

| Dataset | Source | Unique contents | Resolution | Frame rate | Duration (sec) |
|---|---|---|---|---|---|
| KoNViD-1K [31, 32] | YFCC100M [33] | 1,200 | 540p | 24-30 | 8 |
| KoNViD-150K [29] | YFCC100M [33] | 153,841 | 540p | 24-30 | 5 |
| Live-VQC [34] | From authors | 585 | 240p-1080p | 19-30 | 10 |
| YouTube-UGC [30] | YouTube | 1,500 | 360p-4K | 15-60 | 20 |
| CVD2014 [35] | From authors | 5 | 480p,720p | 9-30 | 10-25 |
| Maxwell [36] | YFCC100M [33] & action datasets [37, 38] | 4,543 | 540p | 24-30 | 8 |
| LIVE-Qualcomm [39] | From authors | 208 | 1080p | 30 | 15 |
| LIVE-FB LSVQ [40] | Meta | 39,000 | 144p-8K | various | 5-12 |
| Youku-V1K [12] | YouKu | 1072 | 540p-1080p | 25-60 | 10 |

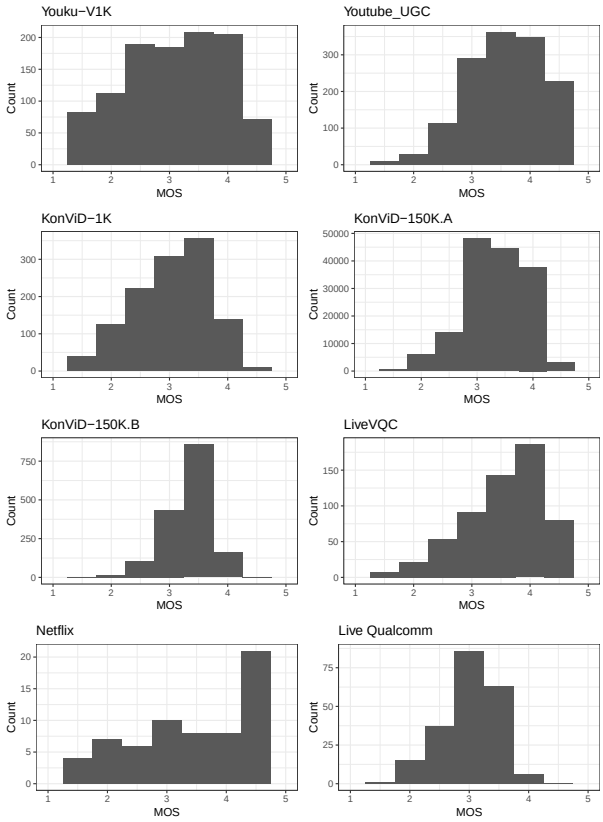


Figure 1: Distribution of subjective rating of openly available video quality datasets

### *Coding condition*

To study the relationship between quality and acceptability of each video, videos were encoded at different quality levels. To define the coding conditions, videos were encoded using the VP9 codec with the preset medium from FFMPEG 6.1. Constant rate factor (CRF) of 10, 20, 28, 32, 36, 40, 45, 50 were used for each considered resolution (1920x1080, 1280x720, 854x480, 640x260, 426x240, 256x144). Quality of each encode is then measured using VMAF by comparing the ingest video in 1920x1080 resolution with each of the encodes. Based on this encodes, the convex hull that provide the highest quality that can be achieved across all resolution for a given bitrate can be estimated. Figure 3 left) depicts this process for one of the video.

VMAF computes the quality of the video with respect to the reference, and relates to DMOS. In order to take into account the quality of the video at ingest, the subjective quality scores that

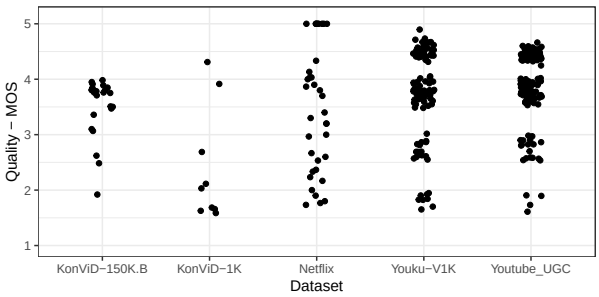


Figure 2: Distribution of subjective rating of ingested videos

were provided from the source dataset are taken into account so to obtain an estimate of the quality of the video after transcoding. Such process is given in Equation 1. This model was based on knowledge extracted from previous experiments, and its validity will be further shown in the result section.

$$MOS_{coded} = 4 \times \frac{VMAF}{100} \times \frac{MOS_{ingest} - 1}{4} + 1 \tag{1}$$

Based on this estimate of the transcoding quality, coding conditions are defined. First, the highest quality is selected so to include the ingest quality. Then, up to 3 other coding conditions were defined. The number of coding condition was set to be dependent on the quality of the ingest. If the ingested video had a low quality, for example a $MOS = 2$, transcoding this video to three other quality level would result in 3 similar quality values which would not be a good use of the experimental resources. Therefore, the process for selecting coding conditions followed these rules:

- The lowest was set to 240p as 144p would lead to too low quality and is expected to be unacceptable.
- The lowest quality 240p video was selected if its quality is distinct enough from the ingest quality.
- Coding conditions are chosen to be uniformly spread in the MOS domain between the highest and lowest quality level.
- There should be at least 0.8 unit of MOS between two coding conditions. If not achievable, fewer transcoding conditions are used.
- Processed video (PVS) were selected among the encoding trials that were performed during the estimation of the convex hull. Videos that most closely match the selection criteria were chosen.

Figure 3 right) depicts the selected coding conditions based on the estimated MOS for these conditions (using Equation 1).

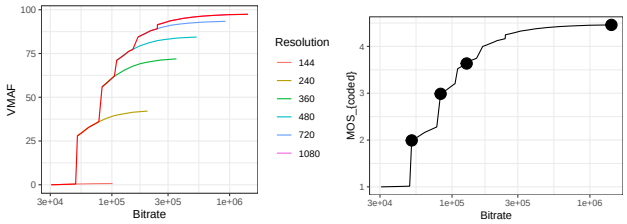


Figure 3: Estimation of the convex hull (left). The convex hull is highlighted with a red curve. On the right side picture, the selection of coding condition based on convex hull and estimation of coded MOS is depicted. The point represents the coding condition that are chosen.

This process is then repeated for every ingested video. Figure 4 depicts the quality distribution of all coding conditions for every ingested video (source content, SRC). In this figure, the vertical axis represents the estimated MOS while the horizontal axis are the different ingested videos (SRCs). Coding condition for the same SRC are linked by a line enabling to visualize the quality range. This figure shows that quality of videos were by design uniformly distributed in the MOS domain.

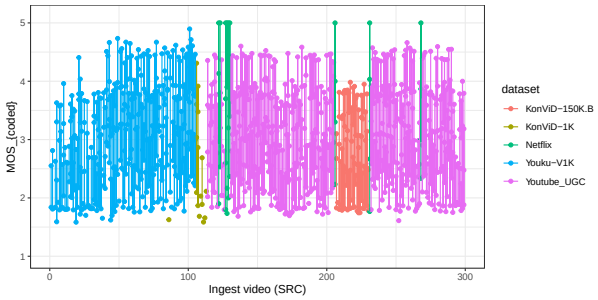


Figure 4: Distribution of estimated MOS across all coding conditions

### *Subjective evaluation*

At the end of the encoding process, 1080 Processed Video Signal (PVSs) needed to be evaluated both in terms of quality and acceptability/annoyance scales. Videos were evaluated in crowdsourcing using an in-house crowdsourcing platform while participants were recruited on Prolific[1].

Considering that participants cannot evaluate all the 1080 videos, 27 playlists of 40 videos were created. Each playlist was constructed so to ensure that participant will see the same range of quality and will also see videos from each of the dataset that were merged by this experiment.

To allow a smooth playback, videos were transcoded into H.264 with the "slower" preset and a CRF of 4. This was done so to preserve quality and ensure that participant will benefit from hardware decoding during video playback. In order to control for the upscaling algorithm that is used for videos with resolution lower than 1920x1080, videos were also pre-upscaled to full HD using the lanczos3 algorithm from FFmpeg. Each of the 27 playlist resulted in approximately 5GB of videos to be downloaded. The downloading process was performed before the experiment could start so to avoid any stalling due to rebuffering issues.

Participants were recruited from western europe and English speaking countries in order to ensure that they would understand the instructions. Screening of the participant was performed to ensure that they would use a computer screen with a resolution of at least HD so they will be able to perceive the full resolution of the videos.

When taking the task, the experiment started with preliminary screening that performed checks on the participants configuration: screen resolution, supported browser, supported operating system, supported devices (only computer were allowed to take the task, phone and tablet users could not take part into the experiment). Then, the participant would be presented with the instruction about the experiment. During the time when participants read the instruction, videos were dowloaded onto the participants' computer. Once the download was completed, participants could start the experiment.

The experiment started with a training procedure where 5 videos of various quality were presented to them. Instruction were given to the participants so to explain the expected rating for these video. Once the participants completed the training, they were able to start the main experiment. To monitor for data quality during the test, honey pot videos were included into the playlists using conditions with clear ratings such as reference video from the Netflix dataset or obviously low quality videos.

In the case of the quality evaluation task, the Absolute Category Rating (ACR) 5 point methodology was used with the label as defined in ITU-T P.910 [41]. In the case of the acceptability/annoyance (AccAnn) evaluation task, participants were asked to rate videos on three categorial item: "not annoying", "acceptable but annoying", "not acceptable" as defined in [11].

In total, 966 participants took part in the quality evaluation task for the 27 playlists, and so far 112 participants took part in the AccAnn evaluation task for the evaluation of 4 of the 27 playlists. In the quality evaluation task, videos received between 32 and 41 ratings. In the case of the AccAnn experiments, each of the video from the completed playlists received from 26 to 50 ratings.

## Results

In this section, the results of the crowdsourcing experiments are presented.

### *Evaluation of subjective quality*

Once data were collected, quality scores were processed using the sureal package[2]. This provides an implementation of a model of subject bias/inconsistency and maximum likelihood estimation (MLE). This model is part of the ITU-T Recommendation P.910 [41]. Using this model, inconsistencies from participant can be taken into account and allow recovering MOS without the contribution from participants with strong inconsistencies.

Figure 5 depicts the relationship between MOS values collected in this crowdsourcing experiment compared to the MOS data from the respective original datasets. Table 2 depicts the relationship between the quality ratings from the conducted crowdsourcing experiment and the ratings provided in the original dataset. The table shows that, with the exception of KonViD-1K, strong correlation can be found between the collected data

[1] https://www.prolific.com/

[2] https://github.com/Netflix/sureal

in this experiment and the data from the open datasets. Equation 2 describe a linear model between the data collected in this crowdsourcing experiment ($MOS_{cs}$) and the data from the original datasets ($MOS_{ds}$). From the regression coefficients, it can be observed that the videos from YouTube UGC and KonViD-150K.B datasets were rated at lower quality in this experiment compared to the ratings in their original dataset. Such difference show the importance of context in the evaluation and participants rate with respect to the other video they see when doing their task.

Table 2: Analysis of the alignment between between the different datasets and the conducted crowdsourcing experiment. PCC refers to the pearson linear correlation coefficient, and SRCC is the spearman rank order coefficient. S and O are respectively the slope and offset in Equation 2.

| dataset | S | O | PCC | SRCC |
|---|---|---|---|---|
| Youku-V1K | 0.826 | 0.542 | 0.928 | 0.862 |
| KonViD-1K | 1.038 | -0.308 | 0.780 | 0.733 |
| Youtube_UGC | 0.639 | 1.496 | 0.932 | 0.879 |
| Netflix | 0.893 | 0.551 | 0.957 | 0.958 |
| KonViD-150K.B | 0.545 | 1.618 | 0.926 | 0.889 |

$$MOS_{ds} = S \times MOS_{cs} + O \tag{2}$$

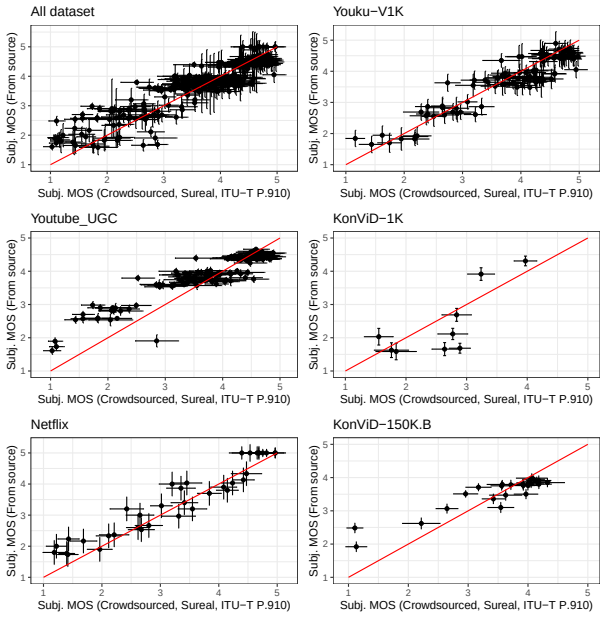


Figure 5: Alignment between the crowdsourced experiments conducted in this study and the respective original datasets.

Figure 6 and 7 depicts the relationship between subjective quality scores obtained from the crowdsourcing experiment, and the prediction from the model described in Equation 1, and the predictions from the UVQ compression distortion network [16]. The model from Equation 1 that combines the subjective score from the origin subjective dataset with VMAF scores shows a strong correlation with the crowdsourced quality ratings (PCC of 0.920, SRCC of 0.919). This shows that combining VMAF with ingest MOS in a multiplicative manner as shown in Equation 1 allows predicting the quality of transcoded videos. However, the quality of the ingested videos need to be known. In this case, MOS data from the source database were used, but such MOS value is usually unknown. The UVQ compression distortion network is a no-reference video quality model and can predict quality in a blind manner. Performance numbers show a PCC of 0.835 and a SRCC of 0.839. Finally, Figure 8 depicts the combined use of UVQ and VMAF as in Equation 1. With this approach, ingest video quality is estimated using UVQ, and degradation compared to ingested quality is measured using VMAF. Both measures are then combined multiplicatively. With this approach, quality prediction accuracy reaches a PCC of 0.867 and a SRCC of 0.869 showing a slight improvement of performance.

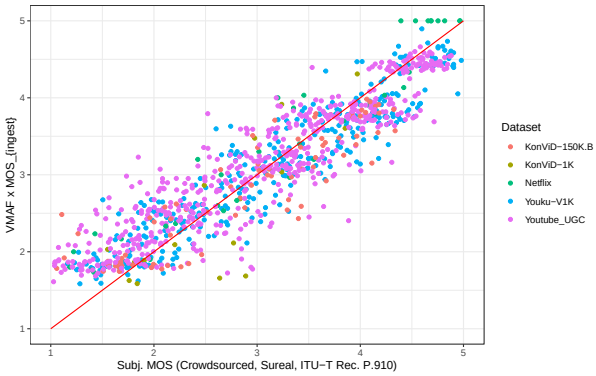


Figure 6: Relationship between predicted MOS using Equation 1 with subjective scores from the crowdsourced experiment.

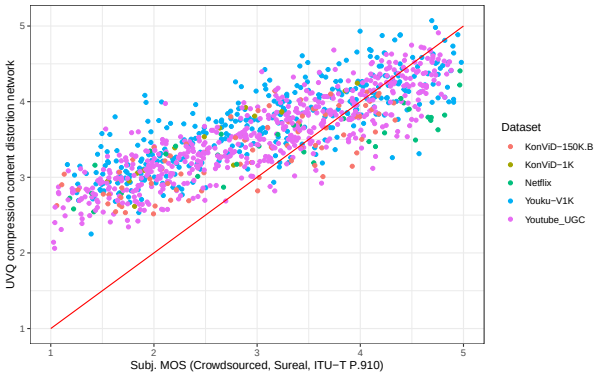


Figure 7: Relationship between predicted MOS from UVQ compression distortion network [16] with subjective scores from the crowdsourced experiment.

### *Study of acceptability/annoyance*

In the second part of the crowdsourcing study, participants rated videos on the acceptability/annoyance scale. The videos are classified into three categorial classes: not acceptable, acceptable but annoying, acceptable and not annoying. Considering the categorial nature of the label, an average rating cannot be computed directly from the individual answers from the participants. Therefore, Li [11] defined an algorithm that make use of Fisher's exact test to identify which AccAnn class should be attributed to each video based on individual observers ratings. Videos are then classified into "not acceptable" (NAcc), "acceptable but annoying" (AA), "not annoying" (NAnn). To account for cases where the Fisher's exact test indicates that statistical significance was not

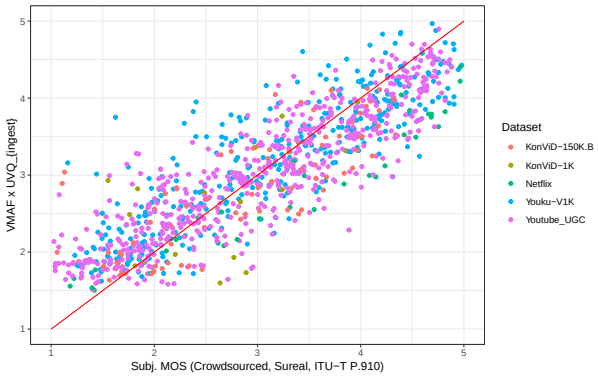


Figure 8: Relationship between predicted MOS from the combined use of UVQ and VMAF with the subjective scores from the crowdsourced experiment. UVQ is used to predict ingest quality, while VMAF predict degradation from coding.

met, two additional classes are added: unsure about the acceptability but for sure about the annoyance (UAcc) and unsure about the annoyance but for sure about the acceptability (UAnn).

Figure 9 depicts the relationship between the quality ratings and the classes from the AccAnn ratings. It can be observed that as quality increases, acceptability levels increases as well. From this figure, it can be seen that a MOS of 2 and 3.5 were found to be clear thresholds between the "not acceptable", "acceptable but annoying", and "not annoying" classes.

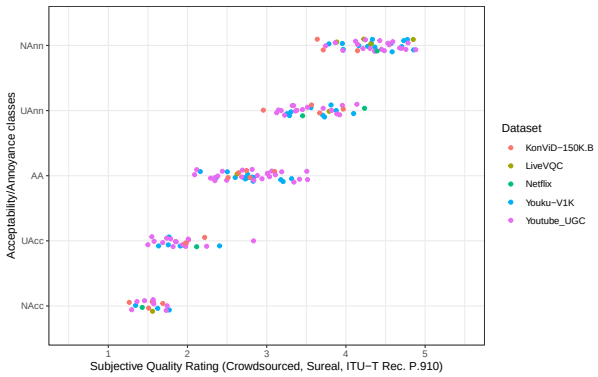


Figure 9: Relationship between subjective quality and AccAnn classes.

Figure 10 depicts the relationship between AccAnn ratings and video quality predicted with UVQ compression distortion network (left), or with UVQ combined with VMAF (right). Results shows that the separation between AccAnn classes becomes more challenging, and further work is needed to predict AccAnn levels of UGC videos.

## Conclusion

In this paper a crowdsourcing experiment was conducted to study the quality and acceptability/annoyance of videos. It was shown that there is a misalignment between existing open UGC datasets, and subjective ratings from one dataset are not in the same quality scale than the other datasets. The performance of different quality estimation model was tested. It was shown that if ingest video quality is known, VMAF is capable of predicting the quality of transcode of these ingest video by using a simple mul-

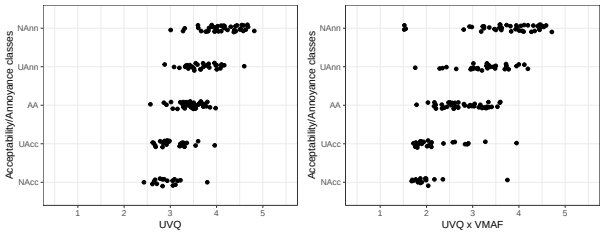


Figure 10: Relationship between quality from prediction model and AccAnn classes.

tiplication between ingest quality and VMAF scores. It was also shown the performance of UVQ at estimating transcoded video quality could be improved by combining it with VMAF. The acceptability/annoyance of videos were evaluated, and it was found that a simple threshold of 2 and 3.5 could distinguish between "not acceptable", "acceptable but annoying", and "not annoying" -quality of videos. Finally, performance of UGC quality prediction model at separating AccAnn levels was tested, but shows space for further improvement. In future work, the modeling of the likelihood of each AccAnn classes as a function of MOS will be studied. Furthermore, we will work on improving the prediction accuracy of AccAnn levels.

## 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*, 2021.

[2] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," in *The Netflix Tech Blog*, 2016.

[3] Neflix, "Per-title encode optimization," in *netflix technology blog*, 2015.

[4] Megha Manohara, Anush Moorthy, Jan De Cock, Ioannis Katsavounidis, and Anne Aaron, "Optimized shot-based encodes: Now streaming!," in *The Netflix Tech Blog*, 2018.

[5] Anush Krishna Moorthy and Alan Conrad Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010.

[6] Peng Ye, Jayant Kumar, Le Kang, and David Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[7] 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.

[8] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network.," in *ECCV*, 2018.

[9] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *26th ACM Int. Conf. Multimedia*, 2018, pp. 546–554.

[10] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[11] Jing Li, Lukáš Krasula, Yoann Baveye, Zhi Li, and Patrick Le Callet, "AccAnn: A new subjective assessment methodology for measuring

acceptability and annoyance of quality of experience," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2589–2602, 2019.

[12] Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen, "Perceptual quality assessment of internet videos," in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA, 2021, MM '21, p. 1248–1257, Association for Computing Machinery.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[14] Junyong You and Jari Korhonen, "Transformer for image quality assessment," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1389–1393.

[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang, "MUSIQ: Multi-scale image quality transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5128–5137.

[16] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang, "Rich features for perceptual quality assessment of UGC videos," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13430–13439.

[17] S. Jumisko-Pyykko, V. K. M. Vadakital, and M. M. Hannuksela, "Acceptance threshold: A bidimensional research method for useroriented quality evaluation studies," in *International Journal of Digital Multimedia Broadcasting*, 2008.

[18] ITU-T Recommendation P.10/G.100, "Vocabulary for performance, quality of service and quality of experience," in *ITU-T*, 2017.

[19] W. Song and D. W. Tjondronegoro, "Acceptability-based QoE models for mobile video," *IEEE Trans. on Multimedia*, vol. 16, no. 3, 2014.

[20] Pierre Lebreton and Kazuhisa Yamagishi, "Study on viewing completion ratio of video streaming," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[21] Pierre Lebreton and Kazuhisa Yamagishi, "Predicting user quitting ratio in adaptive bitrate video streaming," *IEEE Transactions on Multimedia*, vol. 23, pp. 4526–4540, 2021.

[22] M. Sasse and H. Knoche, "Quality in context - an ecological approach to assessing QoS for mobile TV," in *Proc. 2nd ISCA/DEGA Workshop Perceptual Quality System*, 2006.

[23] P. Spachos, W. Li, M. Chignell, L. Zucherman, and J. Jiang, "Acceptability and quality of experience in over the top video," in *IEEE ICC 2015 - Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI)*, 2015.

[24] R. Apteker, J. Fisher, V. Kisimov, and H. Neishlos, "Video acceptability and frame rate," *IEEE Multimedia*, vol. 2, no. 3, pp. 32 – 40, 1995.

[25] T. C. M. de Koning, P. Veldhoven, H. Knoche, and R. E. Kooij, "Of MOS and men: Bridging the gap between objective and subjective quality measurements in mobile TV," in *Multimedia on Mobile Devices IS&T/SPIE Symposium on Electronic Imaging*, 2007.

[26] ITU-T Recommendation J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," in *ITU-T*, 2004.

[27] T. D. Pessemier, K. D. Moor, A. J. Verdejo, D. V. Deursen, W. Joseph, L. D. Marez, L. Martens, and R. V. de Walle, "Exploring the acceptability of audiovisual quality for mobile video session based on objectively measured parameters," in *Quality of Multimedia Experience*, 2011.

[28] Ali Ak, Anne Flore Perrin, Denise Noyes, Ioannis Katsavounidis, and Patrick Le Callet, "Video consumption in context: Influence of data plan consumption on QoE," in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, New York, NY, USA, 2023, IMX '23, p. 320–324, Association for Computing Machinery.

[29] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe, "KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild," *IEEE Access*, vol. 9, pp. 72139–72160, 2021.

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

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

[32] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, "The konstanz natural video database (KoNViD-1k)," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[33] B. Thomee, D.A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li, "YFCC100M: The new data in multimedia research," in *Communications of the ACM*, 2016, vol. 59, pp. 64–73.

[34] Zeina Sinno and Alan Conrad Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2019.

[35] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen, "CVD2014-a database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, 2016.

[36] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin, "Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach," in *31st ACM International Conference on Multimedia (MM '23)*, 2023.

[37] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman, "The kinetics human action video dataset," in *arXiv:1705.06950*, 2017.

[38] Chunhui Gu, Chen Sun, and et al, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *CVPR*, 2018.

[39] Deepti Ghadiyaram, Janice Pan, Alan C. Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2061–2077, 2018.

[40] Z. Ying, M. Mandal, D. Ghadiyaram, and A.C. Bovik, "Patch-VQ: 'patching up' the video quality problem," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14019–14029.

[41] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," in *ITU-T*, 2023.

## Author Biography

*Pierre David received the engineering degree from the Institut d'Optique Graduate School—SupOptique, Palaiseau, France, and the M.Sc. degree in signal processing from the University of Bordeaux, Bordeaux, France, both in 2016. He received a Ph.D. degree in signal and image processing with the University of Rennes 1, at INRIA, Rennes, France. His research interests include light field imaging, motion estimation, and frame interpolation. He is currently working at CAPACITES, Nantes, France where he focuses on transferring research to businesses.*

*Pierre Lebreton received the Engineering degree in computer science from Polytech' Nantes, Nantes, France, in 2009. In 2010, he joined the Group Assessment of IP-based Applications, Berlin Institute of Technology, Berlin, Germany, where he studied toward his Ph.D. on 3D video QoE. After graduating, he joined the Group of Audio Visual Technology, TU-Ilmenau, Germany, in 2015, and the Group of Networked Sensing and Control, Zhejiang University, Hangzhou, China, in 2016. His research interests include various topics including aesthetic appeal, large scale video quality monitoring, and bike sharing systems. In 2017, he joined NTT Laboratories, where he focused on quality, user-engagement prediction and techniques for encoding and streaming videos. Since June 2023 he works at CAPACITES, Nantes, France where he focuses on transferring research to businesses.*

*Neil Birkbeck received the Ph.D. degree from the University of Alberta in 2011 working on topics in computer vision, graphics and robotics, with a specific focus on image-based modeling and rendering. He went on to become a Research Scientist with Siemens Corporate Research working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Software Engineer with the Media Algorithms Team, YouTube/Google. His research interests include perceptual video processing, video coding, and video quality assessment.*

*Yilin Wang received the B.S. and M.S. degrees in computer science from Nanjing University, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from the University of North Carolina at Chapel Hill in 2014, working on topics in computer vision and image processing. After graduation, he joined the Media Algorithm Team, Youtube/Google. His research interests include video processing infrastructure, video quality assessment, and video compression.*

*Balu Adsumilli received the master's degree from the University of Wisconsin-Madison in 2002 and the Ph.D. degree from the University of California at Santa Barbara in 2005, working on watermark-based error resilience in video communications. He currently manages and leads the Media Algorithms Group, YouTube/Google. From 2005 to 2011, he was a Senior Research Scientist with Citrix Online and from 2011 to 2016, he was a Senior Manager Advanced Technology with GoPro, at both places developing algorithms for images/video quality enhancement, compression, capture, and streaming. He is an active member of IEEE (and MMSP TC), ACM, SPIE, and VES. He has coauthored more than 120 papers and patents. His research interests include image/video processing, machine vision, video compression, spherical capture, VR/AR, visual effects, and related areas.*

*Patrick Le Callet (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in image processing from the Ecole Polytechnique de 1'Université de Nantes. He was an Assistant Professor from 1997 to 1999 and a full-time Lecturer with the Department of Electrical Engineering, Technical Institute of the University of Nantes, from 1999 to 2003. He led the Image and Video Communication Laboratory, CNRS IRCCyN, from 2006 to 2016, and was one of the five members of the Steering Board of CNRS, from 2013 to 2016. Since 2015, he has been the Scientific Director of the Cluster Ouest Industries Cratives, a five-year program gathering over ten institutions (including three universities). Since 2017, he has been one of the seven members of the Steering Board of the CNRS LS2N Laboratory (450 researchers), as a Representative of Polytech Nantes. He is mostly involved in research dealing with the application of human vision modeling in image and video processing.*