https://doi.org/10.2352/EI.2024.36.11.HVEI-217

Comparison of subjective methodologies for local perception of distortion in videos and impact on objective metrics resolving power

Andréas Pastor[1], Lukáš Krasula[2], Xiaoqing Zhu[2], Zhi Li[2], Patrick Le Callet[1, 3]
[1]Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[2]Netflix Inc., Los Gatos, CA, USA [3]Institut universitaire de France (IUF)

Abstract

Different subjective methodologies exist to collect data on human perception of distortions, from rating methodologies with single or double stimuli to ranking with pairwise comparisons. The Maximum Likelihood Difference Scaling (MLDS) method uses triplet/quadruplet-based comparisons as a ranking task. Participants compare intervals inside pairs of stimuli: (a,b) and (c,d). The task is to rank if they perceive greater differences between (a,b) or (c,d). From these comparisons' judgments, we can place the assessed stimuli on a perceptual scale (e.g., from low to high quality) with the help of a mathematical solver.

However, one limitation is that the perceptual scales retrieved from stimuli of multiple contents are usually different. We previously offered a solution to measure the inter-content scale of multiple contents. In this work, we compare multiple rating and ranking methodologies. We examine obtained subjective quality scores regarding precision by analyzing discriminability in the scores, efficiency by comparing fixed experimental effort costs, and robustness of retrieve estimates to outliers and spammer behaviors. In this work, we put data quality, experimental cost, and resolving power into relation. We show how discriminability in the data impacts the resolving power of popular objective quality metrics. Our findings are that higher-performing metrics require higher-quality data to reveal their full potential.

Introduction

Subjective quality assessment methodologies provide essential feedback on the quality of a system and how users perceive it. They are necessary to benchmark objective quality metrics and to create datasets to train machine learning and deep learning models. However, running an in-lab or crowdsourced experiment takes time and effort. Furthermore, due to the subjectivity of the stimuli and the task of annotating them, there is not always an agreement in people's judgment. Accurate estimations are needed to reduce noise and uncertainty in collected data. It is then critical to select the most suited methodology to boost and allocate the annotation resources to the proper set of stimuli.

Multiple methodologies exist to rate stimuli, with absolute or relative quality estimations like those performed by Absolute Category Rating (ACR) or Degradation Category Rating (DCR) protocols defined in ITU standard [1]. Moreover, methods based on a ranking of stimuli are two–Alternative Forced Choice (2–AFC), pairwise comparison (PC), and protocols involving triplets or quadruplets. Ranking methodologies are more sensitive than ACR or DCR since observers only need to provide their preference over a set of stimuli, which makes it easier to build a judgment, reduce cognitive load, and thus increase the sensitivity of a subjective experiment. This improves what we call the *discriminability* between stimuli.

Ranking-based subjective experiments yield matrices of choices indicating how often a stimulus is preferred over another. A Pair Comparison Matrix (PCM) requires a model and mathematical solver to translate to a continuous scale. Examples of models are Thurstone [2], Bradley and Terry [3], and Tversky [4]). Due to the pairwise manner of presenting stimuli, the PCM size and the number of possible comparisons rise quadratically with the number of stimuli. Hence, the *efficiency* of a subjective protocol becomes essential to reduce the number of comparisons to perform. Multiple previous works focused on *active-sampling* solutions [5, 6, 7, 8, 9] and more recently [10, 11, 12, 13] to select only the most informative pairs and minimize experimental effort while holding accurate estimations and *robustness* to poor annotator behavior (e.g., spammers).

Another ranking-based methodology is the Maximum Likelihood Difference Scaling (MLDS) methodology [14, 15]. It estimates how pre-ordered stimuli are perceived from comparisons of triplets or quadruplets to retrieve supra-threshold perceptual differences. Stimuli are usually generated from a reference stimulus with an increasing alteration process (e.g., encoding, color-grading, rotation). Preferences of observers on the triplets or quadruplets are aggregated in matrices and used to estimate a perceptual scale per evaluated reference stimulus. This method lacks a global scale where all sets of stimuli from multiple reference stimuli can be represented and scaled. [16] presents a solution to modify the MLDS solver to address this limitation. Validation of the solution is performed through simulated annotations. In [17], more validation is performed on actual data from three datasets of annotations using only ranking-based methodologies (i.e., quadruplets, triplets, and pairwise comparisons). In this work, we compare to an additional fourth dataset. This new dataset of subjective opinions was collected using the DCR methodology [1]. All datasets are available on GitHub[1].

Based on these data, we illustrate how subjective test methodologies influence the quality of collected subjective data by focusing on discriminability among estimated stimuli Mean Opinion Scores (MOS) [18, 19, 20, 21]. This aspect is crucial for objective metrics development and standardization activities, as evaluating and comparing system performances is essential. Sta-

[1]https://github.com/andreaspastor/MLDS_inter_content_scaling

tistical testing that considers subjective data reliability is commonly used for comparing differences between correlations: Root Mean Square Error (RMSE) [22], Spearman Rank Order correlation (SRCC) with the usage of MOS Confidence Interval (CI) [23], and methods estimating Resolving Power of objective metrics [24, 25, 26]. However, this is an afterthought, and the subjective data quality requirement and how much one shall invest in subjective testing to allow discriminating between different quality estimation models has only been weakly studied. We focus on popular Video Quality Assessment (VQA) models validated to local perception estimation in [27], such as VMAF [28], to demonstrate how subjective data discriminability affects our ability to compare these metrics. Our study encompasses different test methodologies, offering a unique perspective on the relationship between experimental cost and the conclusions drawn from the data. Our contributions include:

- Comparative analysis of subjective quality assessment methodologies for evaluating small video patches as a proxy for local distortion perception in videos.
- Examination of experiment cost and its connection with discriminability.
- Investigation of metric resolving power as influenced by subjective discriminability.

Local perception of distortion in videos

Estimating the perception of local distortions by the Human Visual System (HVS) is paramount in video compression. Since distortions introduced by video encoding algorithms can change the perception of content. Encoding is performed through Rate-Distortion estimation and optimization to achieve the highest quality while reducing the amount of bitrates consumed to store the encoded information. These decisions (e.g., selection of modes, partitionings, transforms) operate at the scale of Coding Units (CU) or Coding Tree Units (CTU) levels, which are spatially located, 128×128 pixels for the largest CU in AV1[2].

On the other hand, the human eye performs gazes and saccades to focus on specific regions of interest. Each gaze is spatially localized, with the highest resolution achieved in the central region of the retina: the fovea region covers 1° of visual angle and has a sensitivity of 30 cycles per degree. Gazes are also temporally located with a duration averaging around 200 to 300ms when performing a fixation. Lastly, eye movements can track moving objects and anticipate their future position, commonly called smooth pursuit. Hence, the connection between these two aspects, video encoding CTU and HVS fovea, is not straightforward and needs to be understood. We define Percpetual Units (PU) as this spatio-temporal granularity of the HVS we want to understand and model to predict the perception of distortions and integrate it in video encoders. Figure 1 illustrates an example of PU. A PU is a 64x64 block of pixels and spans over 12 consecutive frames (400ms).

To understand how the HVS perceives distortions introduced by video encoding algorithms, we must collect data from human observers to possess a subjective ground truth. The choice of subjective methodology to collect, evaluate, and train objective metrics is essential. Multiple methodologies exist to collect the perceived quality of visual stimuli.

[2]AV1 encoder v3.1.2, from AOM Alliance Open Media: https://aomedia.googlesource.com/aom/

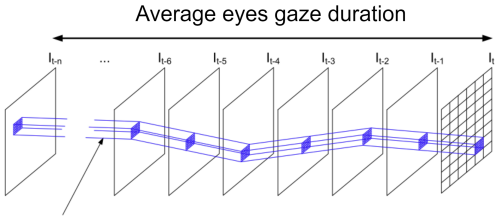

Figure 1. *Example of a Perceptual Unit (PU) aligned on motion. The PU is a 64x64 block of pixels and spans over 12 consecutive frames (400ms). The PU is the spatio-temporal granularity of the Human Visual System (HVS) we want to understand and model to predict the perception of distortions and integrate it into video encoders.*

Each subjective methodology requires a different number of observers and a different number of presentations per stimulus. The choice of the methodology will impact the time, cost, and quality of the data collection. An efficient and reliable methodology is needed to collect data on these Perceptual Units. Moreover, the results of methodologies need to be compared to select the most adapted and efficient one.

As we have PU to define the spatio-temporal granularity of the HVS we want to model, we can also define tube-contents. A tube-content TC_i is a set of tubes, with a reference tube extracted at a specific spatiotemporal location in the video source (SRC) and then multiple versions of this reference tube but distorted with different encoding parameters. Hence, this set can be written as $TC_i = \{TC_{ref}, TC^1_{dist}, ..., TC^N_{dist}\}$. The size of these tubes follows the definition of a PU.

Subjective Datasets on Perceptual Units

This section presents the subjective datasets we collected using quadruplets, triplets, pairs, and double stimuli for local perception of distortions in videos.

The stimuli evaluated are from the dataset presented in [27]. *Tube-contents* are extracted from high-quality videos encoded using AV1, reflecting types of contents and encoding recipes used at Netflix. 8 tube-contents were selected with 5 distortion levels in each tube-content $TC_i = \{TC_{ref}, TC^1_{dist}, ..., TC^5_{dist}\}$.

Quadruplet, Triplet and pair based dataset

This section recaps the three subjective experiments conducted in [17] with methodologies using ranking of stimuli.

Quadruplet-based intra-content dataset

We collected data on the 8 tube-contents described above for the quadruplet dataset. Each tube-content with 6 stimuli yields 15 possible quadruplets. From a $TC_i = \{TC_{ref}, TC^1_{dist}, ..., TC^5_{dist}\}$, quadruplets are $\{(TC_{ref}, TC^1_{dist}, TC^2_{dist}, TC^3_{dist}), (TC_{ref}, TC^1_{dist}, TC^2_{dist}, TC^4_{dist}), ..., (TC^2_{dist}, TC^3_{dist}, TC^4_{dist}, TC^5_{dist})\}$.

We divided these 120 quadruplets from the 8 tube-contents into three playlists of 40 trials. We collected 1800 annotations in total, with 15 participants on each of the playlists.

Triplet-based intra-content dataset

In this dataset, we collected data on same 8 tube-contents using triplets generated following the procedure presented under "The Method of Triads", section 5 of [15]. Here, the 6 stimuli of a TC (i.e., reference + 5 levels of distortions) yield 20 triplets to annotate. We divided the resulting 160 triplets into four playlists of 40 trials each. We collected 1760 annotations with 11 participants

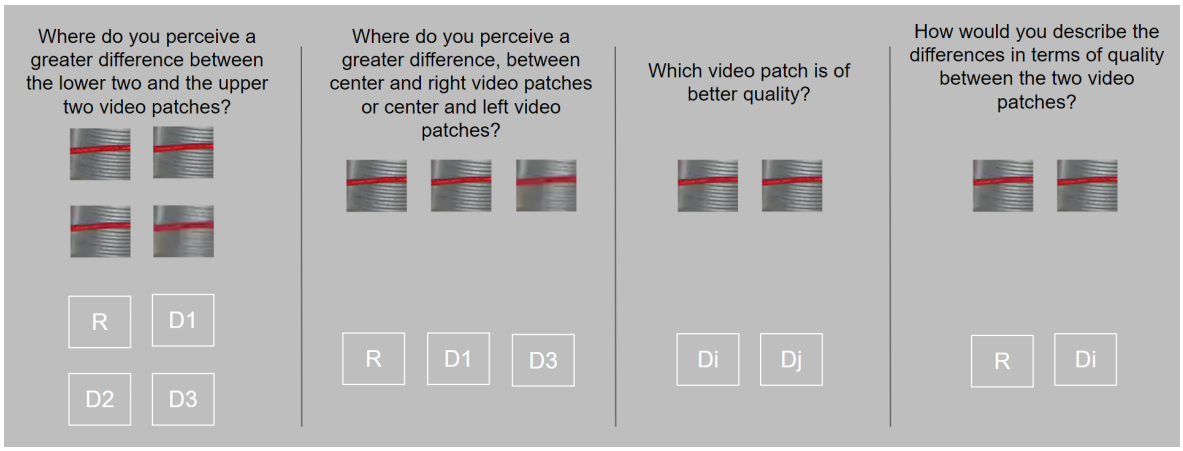

Figure 2. *Displayed are examples of stimuli evaluated by observers in the four subjective tests. Under each example, in white is the information about where the reference (R) and distorted tubes (Di) are located. This information is not provided during tests and is just for the readers. From left to right, the first depicts quadruplets containing four unique distortion levels, while the second showcases triplets comprising three unique distortion levels. These quadruplets and triplets were sampled following the strategy outlined in [14, 15], adhering to the predetermined ordering of stimuli based on distortion strength. The third type of stimuli is pairs for quality ranking. The last type is for the DCR experiment with reference and distorted stimuli pairs. Observers provide an impairment level assessment on the ITU BT.500 impairment scale [1], see Table 1.*

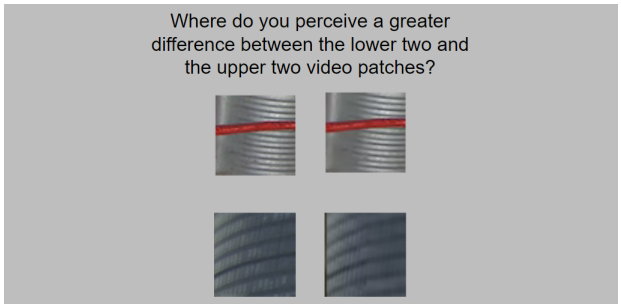

Figure 3. *Example of quadruplet-based inter-content comparison. Each pair in the quadruplet contains the reference tube and a distorted version of this tube. Stimuli are displayed on a neutral gray background.*

on each of the playlists. From $TC_i = \{TC_{ref}, TC^1_{dist}, ..., TC^5_{dist}\}$, triplets are $\{(TC_{ref}, TC^1_{dist}, TC^2_{dist}), (TC_{ref}, TC^1_{dist}, TC^3_{dist}), ..., (TC^3_{dist}, TC^4_{dist}, TC^5_{dist})\}$

Pairwise-based intra-content dataset

This dataset uses pairwise comparisons with 6 stimuli; 15 pairs must be annotated. From $TC_i = \{TC_{ref}, TC^1_{dist}, ..., TC^5_{dist}\}$, pairs are $\{(TC_{ref}, C^1_{dist}), (TC_{ref}, C^2_{dist}), ..., (C^4_{dist}, C^5_{dist})\}$ We decided to divide the 120 pairs into three playlists. We collected 1800 annotations with 15 participants for each playlist.

Quadruplet-based inter-content dataset

We used the active-sampling method proposed in [16] to collect inter-content scaling information. An example of inter-content comparison quadruplet is provided in Figure 3. The data used to initialize the active-sampling is a concatenation of the three previously described datasets to avoid any unfair advantage in later analysis. The sampling procedure was stopped after 55 samplings. Each sampling is a batch that contains 40 quadruplets. In between each sampling, the batch is annotated by an observer and fed to the active sampling algorithm.

DCR methodology based dataset

The test procedure follows the Degradation Category Rating (DCR) method specified in ITU–T Rec. P.910 [29]. Observers use the five-point impairment scale (1: Very annoying, 2: Annoying, 3: Slightly annoying, 4: Perceptible but not annoying, 5: Imperceptible), see Table 1. The first stimulus is the reference tube, displayed on the left, and the second is the impaired tube on the right; see the rightmost example in Figure 2. The average viewing duration is 7 min for the 50 trials in the DCR test. Trials containing twice the reference tube are included in the test as attention checks. We expect observers to vote for "Imperceptible" impairment for these trials. If an observer, for at least one attention check, voted more than "Perceptible but not annoying", all of his answers are discarded for later analysis. 98 observers remain after this cleaning. Moreover, a stabilization phase of four pairs of stimuli is included at the beginning of the test: a no-impairment example, two with clearly perceptible impairment, and one with severely annoying impairment. This phase helps observers adjust to the testing methodology, contents, and interface. Stabilization scores are excluded during later analysis. The stimuli used for calibration differ from those assessed later in the test.

Table 1: 5-grade DSIS scale used in DCR experiment.

Scores	Impairment items
5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Table 2 summarizes all datasets and provides the total number of annotations gathered for each dataset. The term *DCR–equivalent observers* denotes the equivalent number of observers needed to accumulate the same volume of annotations following

Table 2: Summary of datasets collected over the 8 tube-contents with 6 stimuli each and "#" meaning "number of".

Subj. test name	# Annotations	DCR–equivalent Observers
$DCR_{patches}$	4900	98
$Quadruplet_{patches}$	4000	80
$Triplet_{patches}$	3960	79.2
$Pair_{patches}$	4000	80

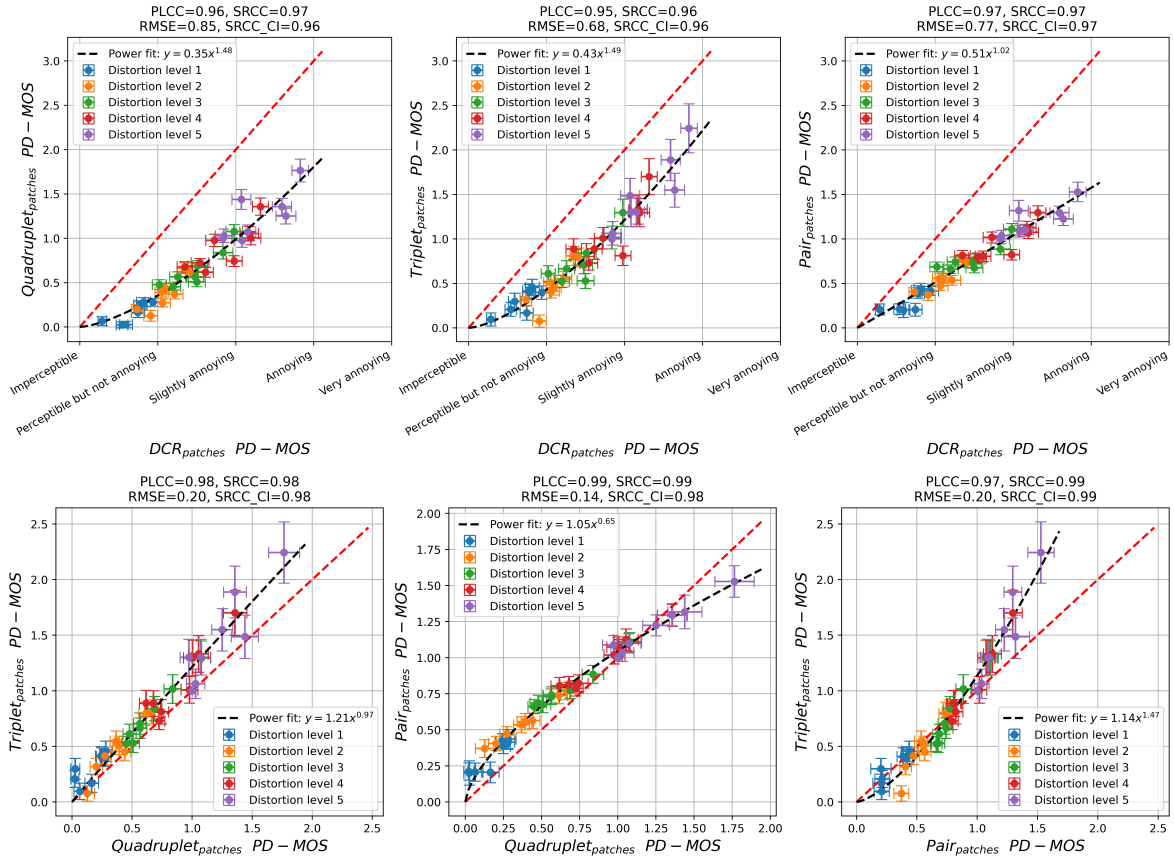

Figure 4. *Comparison of rating scale usage by observers across the four subjective quality assessment methodologies. PD–MOS stands for Perceptual Difference Mean Opinion Score and represents the visibility of distortions perceived by observers. A PD–MOS of 0 corresponds to an imperceptible difference with the reference tube, and increasing PD–MOS represents higher and higher visible distortions.*

the number of trials annotated in a DCR session. For example, the dataset based on quadruplet assessments yielded 4000 annotations, equivalent to 80 DCR–equivalent observers, as calculated by dividing 4000 by 50. 50 is the number of trials annotated in a DCR session.

Comparison of methodologies

This section compares the data collected with the four subjective quality assessment methodologies.

Usage of the scale

Figure 4 analyses the MOS obtained in the four test conditions. We fit a power function (in black) and extract the slope and the power exponent. We analyze these coefficients to see how assessors perceive the stimuli differently and use the rating scales. The red line translates the "one-to-one" relationship where a PD–MOS from one test is equivalent to the PD–MOS of another test. In Figure 4 (top plots), we compare DCR test results with each ranking experiment. We can see strong agreement regarding the Pearson and Spearman correlation (PLCC, SRCC). SRCC_CI from [23] is also inline. Nevertheless, the mapping between PD and MOS is not always linear, translating differences in estimating small distortion ranges. In Figure 4 (bottom plots), PLCC, SRCC, and SRCC_CI agreement is even higher, up to 0.99 correlation. RMSE is smaller, indicating that the PD–MOS score mapping is more consistent across scales and differences in small distortion range are smaller.

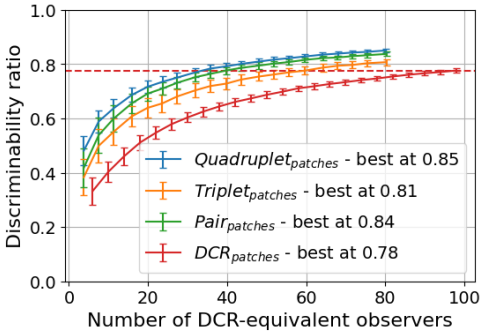

Figure 5. *MOS discriminability evolution for the four datasets, as a function of DCR–equivalent observers' number. Higher discriminability is better.*

Discriminability analysis

In [18, 19, 20, 21], the authors suggested examining the MOS discriminability evolution with increasing numbers of assessors to show how well a subjective methodology can recover accurate MOS scores and compare subjective methodologies efficiency [30]. A two-sample Wilcoxon test is applied to all the possible pairs of MOS in a dataset to test the proportion of significantly different ones. We plot the evolution of this ratio in function of the assessors' number.

Figure 5 presents the results on the four datasets with 95% confidence intervals over 100 simulations. Quadruplet methodology is the most discriminative one. 30 equivalent observers in the

Table 3: Performances of Full-Reference quality metrics on the quadruplet-based dataset.

	PLCC	KRCC	SRCC	RMSE	MAE
VMAF [28]	**0.824**	**0.729**	**0.896**	**0.329**	**0.222**
DLM [31]	0.795	0.684	0.861	0.337	0.223
VIF [32]	0.671	0.669	0.846	0.406	0.302
SSIM [33]	0.742	0.669	0.840	0.384	0.258
PSNR	0.550	0.641	0.803	0.474	0.347

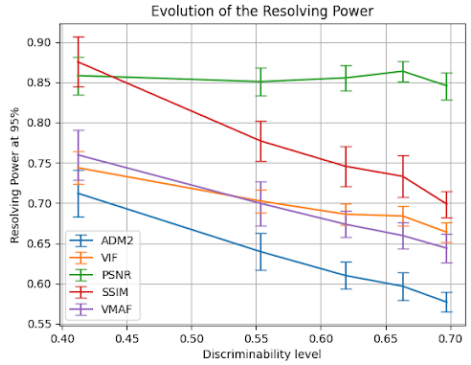

Figure 6. *Objective video quality metrics Resolving Power [25] on the quadruplet-based dataset.*

quadruplet or pair-based scenario achieve the discriminability of the 98 observers in the DCR experiment.

After 40 DCR–equivalent observers (i.e., 2000 annotations), we can comment that for the three ranking-based datasets, continuing to annotate the contents provides a marginal gain in discriminability and thus in data quality. A plateau appears, and adding more time or resources is not necessarily worth it.

Objective Metrics Resolving power

ITU-T Rec. J.149 [25], ITU-R Rec. BT.1676 [24] and in [26] defines how to measure a metric resolving power. Resolving power represents the lowest quality difference a metric can measure that has a statistical difference from the subjective ratings point of view. It characterizes the meaningfulness of quality differences measured by prediction models. However, resolving power depends on metric accuracy and the dataset quality used for evaluating the metric performance. This section will show how much performance the different models can express depending on subjective data quality. Here, we consider five video quality metrics: VMAF [28], DLM [31], VIF [32], MS–SSIM [33], and PSNR. Table 3 presents the performances of these metrics on the quadruplet-based dataset. We chose to analyze this dataset since it achieves the highest discriminability. This analysis could be replicated for the other datasets as well.

Results are presented in Figure 6. Increasing the discriminability, hence the quality of subjective data, improves the resolving power of quality metrics and shows the importance of having reliable PD–MOS. Moreover, linking with results presented in table 3, VMAF performs the best across the different metrics. Its low resolving power also suggests it. Lower resolving power is best. Poorly performing metrics like PSNR have accordingly high resolving power, which does not change much with higher discriminative subjective data.

Conclusion

In conclusion, our study compares multiple subjective methodologies for assessing local perception of distortion in videos and their impact on objective metrics' resolving power. We have shown that ranking-based methodologies, such as quadruplets, triplets, and pairwise comparisons, offer higher discriminability and efficiency than directly rating impairment, like with the DCR method. Quadruplet-based assessments demonstrated the highest discriminability among the tested methodologies, allowing for accurate estimation of perceptual differences with fewer observers than other methodologies.

Furthermore, we illustrated how the quality of subjective data, particularly discriminability among PD–MOS, influences the resolving power of objective quality metrics. Higher discriminability in subjective data leads to improved resolving power, enabling more accurate measurement of quality differences by prediction models. Our analysis highlights the importance of investing in high-quality subjective data collection to leverage objective quality metrics' capabilities fully.

Moving forward, our work lays the groundwork for future research to refine subjective testing methodologies and improve the evaluation of objective quality metrics by better understanding the interplay between subjective data quality, experimental costs, and metric performance. By understanding the relationship between subjective data quality, experimental cost, and resolving power, researchers can make informed decisions to optimize resource allocation and improve the reliability of video quality assessment.

References

[1] ITU Recommendation BT.500-14, "Methodologies for the Subjective Assessment of the Quality of Television Images," International Telecommunication Union, Geneva, 2019.

[2] L. L. Thurstone, "A law of comparative judgement," *Psychological Review*, vol. 34, pp. 278–286, 1927.

[3] R. A. Bradley and M. E. Terry, "The rank analysis of incomplete block designs — I. The method of paired comparisons," *Biometrika*, vol. 39, pp. 324–345, 1952.

[4] A. Tversky, "Elimination by aspects: A theory of choice." *Psychological review*, vol. 79, no. 4, p. 281, 1972.

[5] M. E. Glickman and S. T. Jensen, "Adaptive paired comparison design," *Journal of statistical planning and inference*, vol. 127, no. 1-2, pp. 279–293, 2005.

[6] T. Pfeiffer, X. A. Gao, Y. Chen, A. Mao, and D. G. Rand, "Adaptive polling for information aggregation," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[7] J. Li, M. Barkowsky, and P. Le Callet, "Analysis and improvement of a paired comparison method in the application of 3dtv subjective experiment," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 629–632.

[8] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowd-sourced setting," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 193–202.

[9] P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4249–4256.

[10] J. Li, R. K. Mantiuk, J. Wang, S. Ling, and P. L. Callet, "Hybrid-mst: A hybrid active sampling strategy for pairwise preference aggregation," *CoRR*, vol. abs/1810.08851, 2018. [Online]. Available: http://arxiv.org/abs/1810.08851

[11] Q. Xu, J. Xiong, X. Chen, Q. Huang, and Y. Yao, "Hodgerank with information maximization for crowdsourced pairwise ranking aggregation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] E. Simpson and I. Gurevych, "Scalable bayesian preference learning for crowds," *Machine Learning*, pp. 1–30, 2020.

[13] A. Mikhailiuk, C. Wilmot, M. Perez-Ortiz, D. Yue, and R. K. Mantiuk, "Active sampling for pairwise comparisons via approximate message passing and information gain maximization," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 2559–2566.

[14] L. T. Maloney and J. N. Yang, "Maximum likelihood difference scaling," *Journal of Vision*, vol. 3, no. 8, pp. 5–5, 2003.

[15] K. Knoblauch, L. T. Maloney *et al.*, "Mlds: Maximum likelihood difference scaling in r," *Journal of Statistical Software*, vol. 25, no. 2, pp. 1–26, 2008.

[16] A. Pastor, L. Krasula, X. Zhu, Z. Li, and P. Le Callet, "Improving maximum likelihood difference scaling method to measure inter content scale," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2045–2049.

[17] A. Pastor and P. Le Callet, "Perceptual annotation of local distortions in videos: Tools and datasets," in *Proceedings of the 14th Conference on ACM Multimedia Systems*, ser. MMSys '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 458–462. [Online]. Available: https://doi.org/10.1145/3587819.3592559

[18] Y. Nehmé, J.-P. Farrugia, F. Dupont, P. L. Callet, and G. Lavoué, "Comparison of subjective methods for quality assessment of 3D graphics in virtual reality," *ACM Transactions on Applied Perception (TAP)*, vol. 18, no. 1, pp. 1–23, 2020.

[19] R. F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, "Perceptual evaluation on audio-visual dataset of 360 content," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2022, pp. 1–6.

[20] A. Pastor and P. Le Callet, "Towards guidelines for subjective haptic quality assessment: A case study on quality assessment of compressed haptic signals," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 1667–1672.

[21] A. Pastor, P. Lebreton, T. Vigier, and P. Le Callet, "Comparison of conditions for omnidirectional video with spatial audio in terms of subjective quality and impacts on objective metrics resolving power," Jan. 2024, working paper or preprint. [Online]. Available: https://hal.science/hal-04243995

[22] ITU-T Rec. P.1401, *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*, International Telecommunication Union Std., 2020.

[23] B. Naderi and S. Möller, "Transformation of mean opinion scores to avoid misleading of ranked based statistical techniques," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–4.

[24] ITU-R Rec. BT.1676, *Methodological framework for specifying accuracy and cross-calibration of video quality metrics*, International Telecommunication Union Std., 2004.

[25] ITU-T Rec. J.149, *Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)*, International Telecommunication Union Std., 2004.

[26] L. Krasula, K. Fliegel, P. Le Callet, and M. 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)*, 2016, pp. 1–6.

[27] A. Pastor, L. Krasula, X. Zhu, Z. Li, and P. L. Callet, "On the accuracy of open video quality metrics for local decision in av1 video codec," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 4013–4017.

[28] Netflix, "Vmaf v0.6.1 model," https://github.com/Netflix/vmaf.

[29] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Geneva, 2008.

[30] A. Pastor, P. David, I. Katsavounidis, L. Krasula, H. Tmar, and P. L. Callet, ""Discriminability-Experimental Cost" tradeoff in subjective video quality assessment of codec: DCR with EVP rating scale versus ACR-HR," Dec. 2023, working paper or preprint. [Online]. Available: https://hal.science/hal-04363990

[31] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[32] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.