Perceptual annotation of local distortions in videos: tools and datasets

Andréas Pastor, Patrick Le Callet

Perceptual annotation of local distortions in videos: tools and datasets

Andréas Pastor
Patrick Le Callet
andreas.pastor@etu.univ-nantes.fr
patrick.le-callet@univ-nantes.fr
Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

ABSTRACT

To assess the quality of multimedia content, create datasets, and train objective quality metrics, one needs to collect subjective opinions from annotators. Different subjective methodologies exist, from direct rating with single or double stimuli to indirect rating with pairwise comparisons. Triplet and quadruplet-based comparisons are a type of indirect rating. From these comparisons and preferences on stimuli, we can place the assessed stimuli on a perceptual scale (e.g., from low to high quality). Maximum Likelihood Difference Scaling (MLDS) solver is one of these algorithms working with triplets and quadruplets. A participant is asked to compare intervals inside pairs of stimuli: (a,b) and (c,d), where a,b,c,d are stimuli forming a quadruplet. However, one limitation is that the perceptual scales retrieved from stimuli of different contents are usually not comparable. We previously offered a solution to measure the inter-content scale of multiple contents. This paper presents an open-source python implementation of the method and demonstrates its use on three datasets collected in an in-lab environment. We compared the accuracy and effectiveness of the method using pairwise, triplet, and quadruplet for intra-content annotations. The code is available here: *https://github.com/andreaspastor/MLDS_inter_content_scaling*.

CCS CONCEPTS

• **Human-centered computing**; • **Applied computing**;

KEYWORDS

Subjective methodology, Perception, Multimedia, Quality

ACM Reference Format:
Andréas Pastor and Patrick Le Callet. 2023. Perceptual annotation of local distortions in videos: tools and datasets. In *Proceedings of the 14th ACM Multimedia Systems Conference (MMSys '23), June 7–10, 2023, Vancouver, BC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3587819.3592559

MMSys '23, June 7–10, 2023, Vancouver, BC, Canada

ACM ISBN 979-8-4007-0148-1/23/06…$15.00
https://doi.org/10.1145/3587819.3592559

1 INTRODUCTION

Subjective methodologies provide essential feedback on the quality of a system and how users perceive them. 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 subjective experiment is time-consuming and expensive. 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 subjective methodology to boost and allocate the annotation resources to the right set of stimuli.

Multiple methodologies exist to rate stimuli, with direct estimation like Absolute Category Rating (ACR) or Double Stimuli Impairment Scale (DSIS). Indirect methods are, for example, two-Alternative Forced Choice (2AFC) or pairwise comparison (PC). PC is more reliable since observers only need to provide their preference on a pair, and comparisons are more sensitive. It improves the *discriminability* between stimuli.

PC experiments generate matrices of values indicating how often a stimulus has been preferred over another. These Pair Comparison matrices (PCM) are transformed to a continuous scale, using models (e.g., Thurstone[18], Bradley and Terry [1], Tversky [19]). Due to the pairwise manner of presenting stimuli, the PCM size and the number of possible comparisons grow quadratically with the number of stimuli, introducing *efficiency* in a subjective protocol. A lot of previous works focused on *active-sampling* solutions [2, 4, 6, 7, 16, 22] and more recently [8, 10, 11, 17, 21] to select the most informative pairs and minimize experimental effort while maintaining accurate estimations and *robustness* to bad annotator behavior (e.g., spammers).

In this paper, we focus on providing an understanding of the code we make open-source on GitHub as an extension to the Maximum Likelihood Difference Scaling (MLDS) methodology [5, 9].

MLDS estimates how pre-ordered stimuli are perceived with comparisons of supra-threshold perceptual differences. The stimuli are generated from a reference stimulus with an increasing alteration process (e.g., encoding, color-grading, rotation). We have a perceptual scale per set of stimuli and lack a global scale where all sets of stimuli can be represented. In [15], we presented an approach to address this limitation and validate it through simulated annotations.

In this work, three datasets of annotations using quadruplets, triplets, or pairwise comparisons are made available. They are collected from participants in an in-lab subjective experiment over

small video patches to estimate the local distortions in videos. We provided an analysis and a comparison of the accuracy of the estimates retrieved from these three experiments.

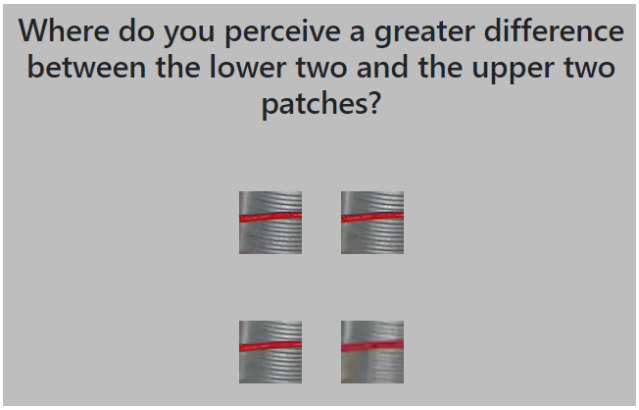

Figure 1: Example of quadruplets that an observer needs to rate. The task is to judge where he perceives a greater difference between the top pair of patches and the bottom pair.

2 MAXIMUM LIKELIHOOD DIFFERENCE SCALING

In this section, we present the MLDS methodology [5, 9] and the extension we introduced in [15] to estimate an inter-content scaling for cross-content perception comparison.

A content C_i has a reference stimulus S_1^i and $n-1$ modified versions of it under test: $S_2^i, ..., S_n^i$. This set of stimuli $C_i = \{S_1^i, S_2^i, ..., S_n^i\}$ is pre-ordered along a physical continuum, with the assumption that larger alterations introduce higher perceptual differences:

$$S_1^i < S_2^i < ... < S_n^i \tag{1}$$

During each trial of the subjective test, a quadruplet (S_i, S_j, S_k, S_l) is presented to the observer; see figure 1 for an example. In return, the observer needs to estimate where he perceives a greater distance between the pair (S_i, S_j) or (S_k, S_l). A trial outcome is 0 or 1, corresponding to the following judgment:

$$\left|S_i - S_j\right| - |S_k - S_l| > 0. \tag{2}$$

During solving, MLDS estimates scalar values $(\phi_1^i, ..., \phi_n^i)$ to fit to the observer judgments.

2.1 Intra-content difference scaling

Judgments of observers can be interpreted as a system of linear equations, with the assumption introduced in the previous section: larger alterations on the reference stimulus present higher perceptual differences. This system can be solved with a General Linear Model (GLM) using a link function $\eta = logit(\pi(x))$ or $probit(\pi(x))$ where $\pi(x)$ is $\mathbb{P}(Y = 1|X_1 = x_1, ..., X_n = x_n)$ and,

$$\pi(x) = F(\phi_1 X_1 + \phi_2 X_2 + ... + \phi_n X_n), \tag{3}$$

where F is the inverse of the link function η. Maximum Likelihood Estimation (MLE) can also be used for solving, more detail in [5, 9]. We provide in our software an implementation of the MLE version and adapt it for inter-content scaling.

We give an example for 5 quadruplets on 7 stimuli (1-7) from a content C_i. Each line describes a quadruplet and stimuli potentially presented to an observer:

$$\begin{pmatrix} 2 & 4 & 5 & 6 \\ 1 & 2 & 3 & 7 \\ 1 & 5 & 6 & 7 \\ 1 & 2 & 4 & 6 \\ 3 & 5 & 6 & 7 \end{pmatrix} \tag{4}$$

yield the following matrix used in the software solver,

$$X = \begin{pmatrix} 0 & 1 & 0 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & -1 & -1 & 1 \\ 1 & -1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 & 1 \end{pmatrix} \tag{5}$$

This matrix X allows estimating the $(\phi_1^i, ..., \phi_n^i)$ for C_i and then be plotted as a perceptual curve. (e.g. each line in fig. 2 corresponds to the ϕ^i of a C_i).

2.2 Inter-contents difference scaling

In the previous section, we introduced the MLDS method and how to estimate perceptual differences from a set of stimuli of a content C_i. However, with this method, the scaling inter-content is not retrieved. In other words, we are missing how stimuli from a content C_i are perceived differently than those from a content C_j.

We added inter-content comparisons to estimate a scaling factor for each perceptual curve (reminder: one perceptual curve per content). Here, a quadruplet is composed of a pair of stimuli from a first content C_i and a pair from a second content C_j: $(S_a^i, S_b^i, S_c^j, S_d^j)$. The observer is asked, similarly as in intra-content comparison, to judge where he perceives a significantly larger perceptual distance:

$$\left|S_a^i - S_b^i\right| - \left|S_c^j - S_d^j\right| > 0 \tag{6}$$

With the addition of this set of inter-content trials, we can construct a larger matrix X, solved with GLM or MLE, and estimate the difference of scales between a group of perceptual curves, refer to figure 2 for a complete example of solving.

$$X = \begin{pmatrix} & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\ & X_1 & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & X_2 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & & \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \end{pmatrix} \tag{7}$$

X is composed of submatrices (X_1, X_2) of intra-content comparisons for a content C_1 and C_2 and a third matrix of inter-content comparisons: the last two rows in the example above.

GLM and MLE implementation for intra and inter-content scaling are available in this software implementation. In addition, we

provided three datasets of inter-content annotations to run a demonstration of the code.

The stimuli evaluated are borrowed from the dataset presented in [14]. We applied MLDS methodology to estimate supra-threshold differences in small videos, *tubes*, encoded using libaom-AV1[1].

In this application context, C_i defined previously is a *tube-content*, a set with a reference tube and 5 levels of distortions: $C_i = \{S_1^i, S_2^i, ..., S_6^i\}$. The reference tube of size 64×64 pixels and 400 ms is extracted from a 1080p video source (SRC). The 5 distorted tubes are extracted at the exact spatial and temporal location but in 5 Processed Video Sequences (PVS). The PVSs are generated from an SRC encoded using libaom-AV1 at various Quantization Parameter (QP) values.

The datasets were collected in an in-lab experiment with 15 subjects. Each session lasts, on average, 7 minutes and contains 40 trials to annotate.

We collected quadruplet, triplet, and pair-based intra-content comparisons. In addition, we collected inter-content comparisons from quadruplets to retrieve the scaling between the tube-contents. It is worth noting that for inter-content comparison, triplets and pairs are not possible with this type of stimuli. Since we want to estimate with a quadruplet (a,b,c,d) how far a stimulus a is from its reference stimulus b, compared with another stimulus c and its reference stimulus d.

2.3 Quadruplet-based intra-content dataset

We collected data on the 8 tube-contents described above for the quadruplet dataset. The 6 stimuli yield 15 possible quadruplets per tube-content. From a $C_i = \{S_1^i, S_2^i, ..., S_6^i\}$, quadruplets are $\{(S_1^i, S_2^i, S_3^i, S_4^i), (S_1^i, S_2^i, S_3^i, S_5^i), ..., (S_3^i, S_4^i, S_5^i, S_6^i)\}$.

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

2.4 Triplet-based intra-content dataset

In this dataset, we collected data on the 8 tube-contents using triplets generated following the procedure presented in the MLDS paper [5]. Here the 6 stimuli (i.e., reference + 5 levels of distortions) yield 20 triplets to annotate for each content. We divided the 160 triplets from 8 tube-contents into four playlists of 40 trials.

We collected 1760 annotations with 11 participants on each of the playlists.

From $C_i = \{S_1^i, S_2^i, ..., S_6^i\}$, triplets are $\{(S_1^i, S_2^i, S_3^i), (S_1^i, S_2^i, S_4^i), ..., (S_4^i, S_5^i, S_6^i)\}$

2.5 Pairwise-based intra-content dataset

This dataset uses pairwise comparisons with 6 stimuli; 15 pairs must be annotated. From $C_i = \{S_1^i, S_2^i, ..., S_6^i\}$, pairs are $\{(S_1^i, S_2^i), (S_1^i, S_3^i), ..., (S_5^i, S_6^i)\}$ We decided to divide the 120 pairs into three playlists. We collected 1800 annotations with 15 participants on each of the playlists.

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

2.6 Quadruplet-based inter-content dataset

We used the active-sampling method proposed in [15] to collect inter-content scaling information. The sampling of the quadruplets was iteratively performed on the estimates recovered by the GLM solver. 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 sessions of 40 quadruplets were annotated.

All the datasets are summarized in table 1.

Type	# playlists	# annotations
Quadruplet intra	3	1800
Triplet intra	4	1760
Pair intra	3	1800
Quadruplet inter	-	2200

Table 1: Summary of datasets collected over the eight tube-contents with 6 stimuli each and "#" meaning "number of".

3 RESULTS

This section will present the estimates obtained from solving for inter-content scaling. The results from the three datasets are available in figure 2.

In figure 2 (left), the MLE solving is applied on a matrix containing the Pairwise-based intra-content dataset and the Quadruplet-based inter-content dataset: in total 4000 annotations. In figure 2 (middle), the same solving is applied to the Triplet-based intra-content dataset with the Quadruplet-based inter-content dataset (e.g., 3940 annotations), and figure 2 (right) the Quadruplet-based intra-content dataset with the Quadruplet-based inter-content dataset (e.g., 4000 annotations). Confidence Intervals (CI) in the three solving are retrieved via bootstrapping [20].

To comment on the results, all the curves start with stimulus 0 being the reference at a perception of 0: The reference being indistinguishable from itself. The stimuli from "patch 1750" seems to have the most visible distortions across the three datasets, with, for example, in 2 (left) solving an estimation of 1.53 on the perception scale for the 6th stimulus. This stimulus corresponds to the highest compression level applied on the patch with libaom-AV1.

In figure 3, we analyzed the impact of increasing the number of annotations for each dataset. Curves always start on the x-axis with annotation count corresponding to 100% of intra-content respective datasets plus 5% of the inter-content comparison dataset. Increments are a fraction of the inter-content comparison dataset from 5% to 100%. The CIs on the curves are obtained via 100 random Monte Carlo sampling simulations.

In figure 3 (left), we compare the evolution of the Root Mean Score Error (RMSE) between the retrieved scores for each stimulus of the dataset and the retrieved scores using 100% of the dataset. Usually, the right part of each curve should be expected to be at 0, but variations can occur due to the bootstrapping involved in the MLE solving. We can see from this figure that the quadruplet and pairwise methodologies converge faster to their final estimates than the triplet-based intra-content method.

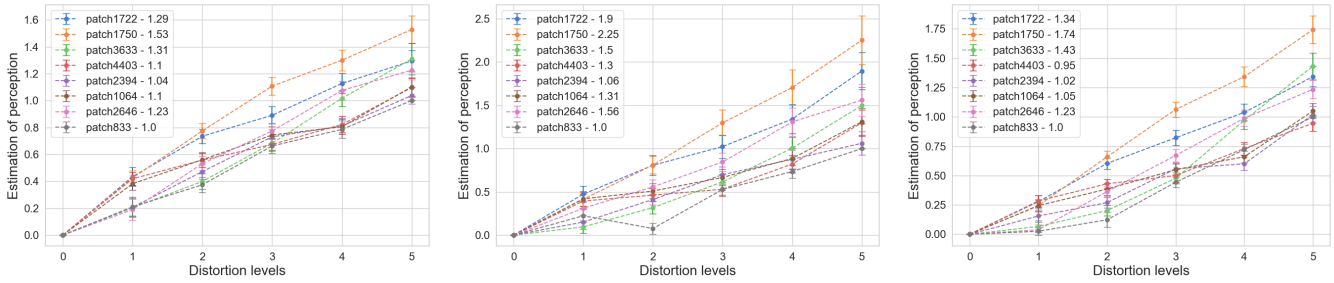

Figure 2: MLE solving on the three datasets with intra-content comparisons using pairs (left), triplets (middle), or quadruplets (right) and also the inter-content quadruplet comparison dataset. CIs are obtained via bootstrapping [20] over 100 runs.

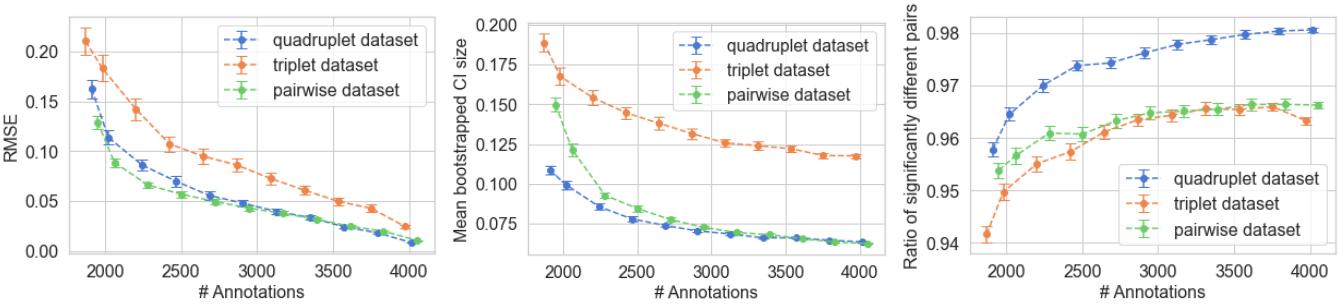

Figure 3: Analysis of the MLE solving accuracy on the three datasets with increasing annotations count (x-axis). On the y-axis, we compare the evolution of RMSE to the estimated ground truth, the size of bootstrapped CI, and the ratio of significantly different pairs in each dataset. CI on the curves obtained via 100 random Monte Carlo sampling simulations.

In figure 3 (middle), the evolution of the size of the CIs estimated by the MLE solver is compared. This value is obtained by taking the average the CI of each score after solving. Again quadruplet and pairwise methodologies' CI sizes are decreasing faster than the triplet-based method. With an advantage for the quadruplet method since, with around 2000 annotations, the average CI size is 0.1 and over 0.125 for the pairwise approach.

Finally, in 3 (right), we analyzed the evolution of discriminability between stimuli with an increasing number of annotations (x-axis). In [3, 12], the authors investigated the growth of subjective scores discriminability to show how well a subjective methodology can retrieve accurate Mean Opinion Scores. A two-sample Wilcoxon test is performed on all the possible pairs of stimuli of the dataset. A p-value of 0.05 is used to compute the percentage of significantly different pairs.

We can see that the quadruplet method is of higher discriminability, with a curve above the triplet and pairwise method.

After 3000 annotations on each dataset, we can comment that adding more inter-content quadruplet annotations does not increase the discriminability or decrease the size of CIs. A plateau is formed. We can conclude that, for our use case of annotation of local distortion in videos, we reach the limit of the methods, and adding more time and resources is not worth it.

4 CONCLUSION

In this work, we provide an open-sourced python implementation of Maximum Likelihood difference Scaling with inter-content difference scaling. We explored a use case to quantify the perceived local distortion in patches of videos encoded using libaom-AV1.

We show from judgments collected, using pairwise, triplet, and quadruplet comparisons, that we can retrieve a scaling inside a group of eight contents. We are proving that our method of inter-quadruplet comparisons can reliably be applied to scale the estimated scores from these three datasets on a unique global perceptual scale.

We analyzed the collected datasets to showcase the accuracy and the discriminability that this method can achieve.

In future work, we plan to collect a large-scale dataset of tube-contents using this method, following the research proposal we presented at ACM MMSys'22 Doctoral Symposium in [13].

ACKNOWLEDGMENTS

This work is supervised by university Professor Patrick Le Callet from Nantes Université, France, and founded by Netflix, Inc. We would like to thanks also people from Netflix for joining our regular meetings and providing valuable feedback: Dr. Lukáš Krasula, Dr. Xiaoqing Zhu, and Dr. Zhi Li.

REFERENCES

[1] Ralph A. Bradley and Milton E. Terry. 1952. The Rank Analysis of Incomplete Block Designs — I. The Method of Paired Comparisons. *Biometrika* 39 (1952), 324–345.

[2] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 193–202.

[3] Randy F Fela, Andréas Pastor, Patrick Le Callet, Nick Zacharov, Toinon Vigier, and Soren Forchhammer. 2022. Perceptual Evaluation on Audio-Visual Dataset of 360 Content. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 1–6. https://doi.org/10.1109/ICMEW56448.2022.9859426

[4] Mark E Glickman and Shane T Jensen. 2005. Adaptive paired comparison design. *Journal of statistical planning and inference* 127, 1-2 (2005), 279–293.

[5] Kenneth Knoblauch, Laurence T Maloney, et al. 2008. MLDS: Maximum likelihood difference scaling in R. *Journal of Statistical Software* 25, 2 (2008), 1–26.

[6] Jing Li, Marcus Barkowsky, and Patrick Le Callet. 2012. 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, 629–632.

[7] Jing Li, Marcus Barkowsky, and Patrick Le Callet. 2013. Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs. In *Stereoscopic Displays and Applications XXIV*, Vol. 8648. International Society for Optics and Photonics, 86481V.

[8] Jing Li, Rafal K. Mantiuk, Junle Wang, Suiyi Ling, and Patrick Le Callet. 2018. Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation. *CoRR* abs/1810.08851 (2018). arXiv:1810.08851 http://arxiv.org/abs/1810.08851

[9] Laurence T Maloney and Joong Nam Yang. 2003. Maximum likelihood difference scaling. *Journal of Vision* 3, 8 (2003), 5–5.

[10] Hui Men, Hanhe Lin, Mohsen Jenadeleh, and Dietmar Saupe. 2021. Subjective Image Quality Assessment With Boosted Triplet Comparisons. *IEEE Access* 9 (2021), 138939–138975.

[11] Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K Mantiuk. 2021. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2559–2566.

[12] Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué. 2020. Comparison of subjective methods for quality assessment of 3D graphics in virtual reality. *ACM Transactions on Applied Perception (TAP)* 18, 1 (2020), 1–23.

[13] Andréas Pastor and Patrick Le Callet. 2022. Perception of Video Quality at a Local Spatio-Temporal Horizon: Research Proposal. In *Proceedings of the 13th ACM Multimedia Systems Conference* (Athlone, Ireland) *(MMSys '22)*. Association for Computing Machinery, New York, NY, USA, 378–382. https://doi.org/10.1145/3524273.3533931

[14] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet. 2022. On the Accuracy of Open Video Quality Metrics for Local Decision in AV1 Video Codec. In *2022 IEEE International Conference on Image Processing (ICIP)*. 4013–4017. https://doi.org/10.1109/ICIP46576.2022.9897469

[15] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet. 2022. 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)*. 2045–2049. https://doi.org/10.1109/ICASSP43922.2022.9746681

[16] Thomas Pfeiffer, Xi Alice Gao, Yiling Chen, Andrew Mao, and David G Rand. 2012. Adaptive polling for information aggregation. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[17] Edwin Simpson and Iryna Gurevych. 2020. Scalable Bayesian preference learning for crowds. *Machine Learning* (2020), 1–30.

[18] Louis Leon Thurstone. 1927. A Law of Comparative Judgement. *Psychological Review* 34 (1927), 278–286.

[19] Amos Tversky. 1972. Elimination by aspects: A theory of choice. *Psychological review* 79, 4 (1972), 281.

[20] Felix A Wichmann and N Jeremy Hill. 2001. The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & psychophysics* 63, 8 (2001), 1314–1329.

[21] Qianqian Xu, Jiechao Xiong, Xi Chen, Qingming Huang, and Yuan Yao. 2018. Hodgerank with information maximization for crowdsourced pairwise ranking aggregation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[22] Peng Ye and David Doermann. 2014. Active sampling for subjective image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4249–4256.