Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in VQA datasets using objective measures

Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in VQA datasets using objective measures

Ahmed Aldahdooh[a,*], Enrico Masala[b], Glenn Van Wallendael[c], Peter Lambert[c], Marcus Barkowsky[d]

[a] *IT Department, University College of Applied Sciences, Gaza, Palestine*
[b] *Control and Computer Engineering Department, Politecnico di Torino, Italy*
[c] *Ghent University - imec - IDLab, Ghent, Belgium*
[d] *Deggendorf Institute of Technology (DIT), University of Applied Sciences, Deggendorf, Germany*

Abstract

Subjective quality assessment is a necessary activity to validate objective measures or to assess the performance of innovative video processing technologies. However, designing and performing comprehensive tests requires expertise and a large effort especially for the execution part. In this work we propose a methodology that, given a set of processed video sequences prepared by video quality experts, attempts to reduce the number of subjective tests by selecting a subset with minimum size which is expected to yield the same conclusions of the larger set. To this aim, we combine information coming from different types of objective quality metrics with clustering and machine learning algorithms that perform the actual selection, therefore reducing the required subjective assessment effort while trying to preserve the variety of content and conditions needed to ensure the validity of the conclusions. Experiments are conducted on one of the largest publicly available subjectively annotated video sequence dataset. As performance criterion, we chose the validation criteria for video quality measurement algorithms established by the International Telecommunication Union.

Keywords: subjective testing, subset selection, statistical analysis, video quality assessment, clustering
2010 MSC: 00-01, 99-00

1. Introduction

The evaluation of the performance gain of innovative video processing technology such as video coding or video post-processing frequently requires assessment by human observers. Objective measures are often used during the development. In production

[*] Corresponding author

Email addresses: adahdooh@ucas.edu.ps (Ahmed Aldahdooh), enrico.masala@polito.it (Enrico Masala), Glenn.VanWallendael@UGent.be (Glenn Van Wallendael), Peter.Lambert@UGent.be (Peter Lambert), marcus.barkowsky@th-deg.de (Marcus Barkowsky)

environments, When used by the industry, for example in broadcast monitoring, measurement algorithms can only be trusted after careful evaluation of their correlation with the results of human scores. Such an evaluation is a tedious task and the outcome is limited to the scope under which the measurement algorithms were evaluated — new video processing technologies come with different types of degradations and therefore require the re-evaluation of the performance for the new use case. The International Telecommunication Union creates new recommendations for industrial use cases on a regular basis, starting with the now somewhat outdated ITU-T Rec. J.144 [1] for evaluating the transmission of standard definition television with the video coding algorithms of 2001 to the latest hybrid video quality measurement algorithms of 2014 in ITU-T Rec. J.343 [2].

The Video Quality Experts Group (VQEG) [3] has conducted the validation experiments for these recommendations. The careful preparation and execution of the validation experiments requires a time frame of two to three years and corresponding effort and cost but also research questions are tackled. Most notably, scientific questions are posed on the selection of video content, the alignment of subjective scores, and the performance comparison of the objective measurement algorithms.

A recurrent research question in these efforts concerns the representativeness of the selected video content (source sequences, SRC) and the degradation procedures and their parameters (hypothetical reference circuits, HRC). Both are currently carefully selected by experts and in most cases a larger set of processed video sequences (PVS) is created from which a subset gets evaluated by human observers. The obtained scores are then compared to the algorithms' results by statistical means and the (set of) best performing algorithm(s) is chosen for recommendation by the ITU under the assumption that the evaluated sequences represent the variety of conditions that the video industry aims to measure. The exact procedure is documented by VQEG in testplan documents and final reports.

A scientific analysis for this kind of testing has been performed by Keimel et.al. in [4]. Alternative approaches have been proposed, in [5] Ciaramello proposes to stress test the algorithms with degradation algorithms that are known to be monotonically decreasing the quality. Based on an analysis of pairwise comparisons with dead-zone, Reibman proposed an optimized selection method for subjective assessments in [6]. An adaptive subjective testing method for the particular case of multiple types of degradations has been developed by Reiter and Korhonen in [7]. In order to find the optimum in this case, Voran and Catellier proposed a gradient ascent strategy that may also be adapted for improving validation experiments [8]. In the reverse case, similar to our approach, employing objective measurement algorithms in conjunction with subjective scores has already helped in the alignment of distinct subjective assessment databases as proposed by Pinson et.al. [9]. In their work, they used objective measures to find sequences with similar quality properties in two (or more) distinct datasets in order to align the subjective votes of these datasets. Their approach allowed to avoid an additional subjective alignment experiment. Our goal is to construct a dataset that minimizes the occurrence of sequences with similar properties in order to reduce the effort of subjective testing.

There are two aspects of the term representativeness which we will later refer to as "aspect 1" and "aspect 2". The first one is that all conditions that may occur in the scope of the actual industrial environment should be taken into consideration, including border cases. The second one is that the algorithms' performance should be optimal on the average use case — including but not overemphasizing the border cases. This calls for a selection of test conditions for subjective assessment by large-scale automatic evaluation prior to the selection by experts.

The purpose of this contribution is to evaluate to which extent clustering with various machine learning algorithms can be used in order to find a representative subset that fulfills both criteria. The aim is to reduce the number of test conditions without changing the results of a validation experiment — conversely, in the next validation experiment, the preselection may help to improve on the representativeness of the test conditions without increasing the number of total scores.

Our contribution starts with an overview of the proposed framework in Section 2, followed in Section 3 by the presentation of the data used for the practical experiments and how they have been derived from the publicly available VQEG-HDTV dataset. Section 4 presents the proposed approach to perform and evaluate the subset selection procedure, followed by results in Section 5. Conclusions are drawn in Section 7.

2. Framework

An overview of the proposed evaluation scheme is shown in Fig. 1. In the first step, n video sequences are collected based on the scope of the evaluation using well-known guidelines, e.g. Pinson et al. [10]. An analysis of the video content with objective characteristics measures ranges from simple Spatial and Temporal perceptual Information measures (SI/TI) specified in ITU-T P.910[11] to more complex signal based characteristics [12] up to the semantical analysis of objects. In the next step, m degradations are added that typically include video coding and image processing algorithms such as color pre- and post-processing, sharpening or image resizing.

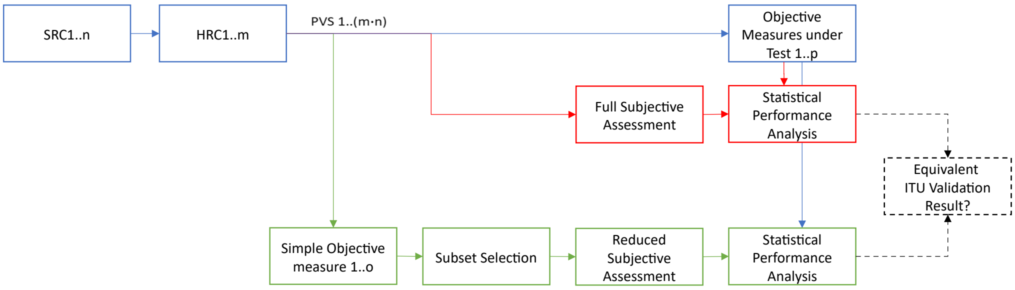
SRC1..n
HRC1..m
PVS 1..(m·n)
Objective Measures under Test 1..p
Full Subjective Assessment
Statistical Performance Analysis
Equivalent ITU Validation Result?
Simple Objective measure 1..o
Subset Selection
Reduced Subjective Assessment
Statistical Performance Analysis

Figure 1: Framework overview

As shown in the figure by the Full Subjective Assessment block, the selected sequences usually get evaluated in one or several subjective experiments. When the subjective assessment is too large to be performed as a whole with a single group of observers, an alignment procedure on the subjective scores should be performed, for example based on a common set of PVS that is shown in each evaluation [13, 14, 15]. Pinson et al. also developed a method using simple objective measures to combine subjective datasets [9], an idea which is similar to our framework. If the merge spans over languages and cultures, special consideration of the perceptual scale attributes is required [16, 17].

In validation experiments, proponents submit their measurement algorithms for independent evaluation prior to the creation of PVS. Once the PVS are established, objective scores or video quality ratings (VQR) are obtained from the proponents' measurement algorithms. It should be noted that, for the sake of unbiased validation, no modification to the original algorithms is allowed although the proponents often improve their commercial products during the timespan in which the PVS are created. This step is shown as Objective Measures under Test in Fig. 1.

Using these subjective and objective scores, a statistical performance analysis is carried out. In this analysis, as more thoroughly described in Sec.4.2, the VQR are mapped to the subjective scale and differences between objective measures are quantified using root mean square error (RMSE) analysis.

Despite the enormous effort that is made, concerns are often raised regarding the validity of the experiment that relate to the representativeness of the chosen PVS as detailed in the introduction.

The overview diagram in Fig. 1 contains our proposal in the last row: First, the quality is evaluated by simple objective measures (1..o). Existing fast measures may be used as the focus is not on precise measurement but only on capturing the variability of the PVS. Examples of measures and how and why they can be used for this purpose will be given in Section 3.2. In the next step, a subset selection is performed that shall reduce redundancies while maintaining the spread of the PVS characteristics. This should lead to a smaller subjective experiment compared to a full evaluation with redundancies, potentially weighting PVS that represent larger or more populated clusters with more observers to get more precise results. Finally, the performance analysis can be done with respect to the coverage of the scope of evaluation (representativeness aspect 1) and with respect to the average behavior (aspect 2).

In this contribution, we focus on the framework's concept by (artificially) reducing the number of PVS that get subjectively evaluated. This limits our analysis to the first aspect because we are missing information about the distribution of the selected PVS in the industrial use case for our dataset. This is left for future work, potentially in an upcoming validation experiment.

3. Dataset

In this work we take advantage of one of the largest subjectively annotated database of high-resolution video sequences, i.e., the so-called Video Quality Expert Group (VQEG) HDTV dataset [18].

In short, such a database has been originally developed in order to evaluate objective video quality models submitted by several proponents. In this paper, these submitted objective quality models will be named model A to F. The quality values stemming from the use of those models have been compared with the results of a rigorous, very well controlled, extensive subjective quality evaluation campaign. Final conclusions were drawn in the VQEG's Final Report approved June 30, 2010 [18].

In addition to two ITU Recommendations [19, 20], such a large effort produced a reliable subjectively annotated database including more than 1,000 PVSs with different types of content and impairments (HRCs). The large majority of such database, made available through the CDVL website [21], can be used for research purposes, e.g., for designing and training new models and algorithms.

VQEG-HDTV dataset consists of six independently evaluated subjective datasets, three of them are progressive (dataset 1,3,5) and can therefore be evaluated by the simple measures as explained in Sec. 3.2.

Note also that several HRCs introduce some temporal misalignments (for example due to transmission errors), which need to be duly considered in case traditional frame-based full-reference (FR) objective models are used. In particular, a temporal registration step is required so that the FR models always compare frames that correspond to the same time instant. This step has been performed by the authors of this work by means of the temporal registration algorithm described in Sec. 3.1. We will only consider FR measures applied to those temporally-registered PVSs.

3.1. Temporal and Spatial Registration

VQEG's testplan [22] defines the limits of spatial, temporal, brightness and color changes that may be found in the PVS due to coding and transmission outages. In order to run image based video quality measurement algorithms, a temporal and spatial registration is required. As we are aiming at measuring all degradations that are present in the PVS, for each frame, a corresponding reference frame is created. In Fig. 2 an overview flow chart is provided. Please note that for each PVS a unique set of two aligned sequences is created.

In this work, we first identify for the complete sequence image blocks that allow for temporal registration, i.e. blocks with high spatial activity. Then, a single global spatial offset is determined based on these blocks, using a temporally subsampled version of the sequence. A horizontal or vertical pixel shift was allowed as a result of hardware encoders or as part of format conversions but it was not allowed to change during the sequence. In this step, each possible spatial offset has to be calculated for each possible temporal shift - the subsampling reduces the computational burden with limited effect on the performance. Next, using the initial blocks, each image is compared to its temporal candidates with the fixed spatial offset. The resulting difference, calculated as Mean Square Error for the blocks, is interpreted as a probability value for the match for each frame. The final temporal registration for each frame is obtained from those probabilities using a maximum likelihood approach similar to the Viterbi algorithm as published in [23].

The exact implementation with all parameters can be found online at: (link will be inserted in the final version of the paper).

3.2. Measures Computation

A number of different measures, referred to as "Simple Objective measure 1..o" in Figure 1, have been employed in our work. The general idea is that using several measures which are sensitive to different aspects of the content can contribute to capture the variability of the PVS. For instance, the widely used Peak Signal-to-Noise Ratio (PSNR) is very sensitive to brightness changes, whereas SSIM [24] and MS-SSIM [25] are particularly sensitive to contrast changes and structural distortion, and VIF [26] takes into more consideration human visual system characteristics. Therefore, we presume that the variability of the PVS is different if such measures give different results.

The publicly available VQMT software [27] is a flexible tool that can compute all these frame-based objective quality measures. For those measure, the VQMT software

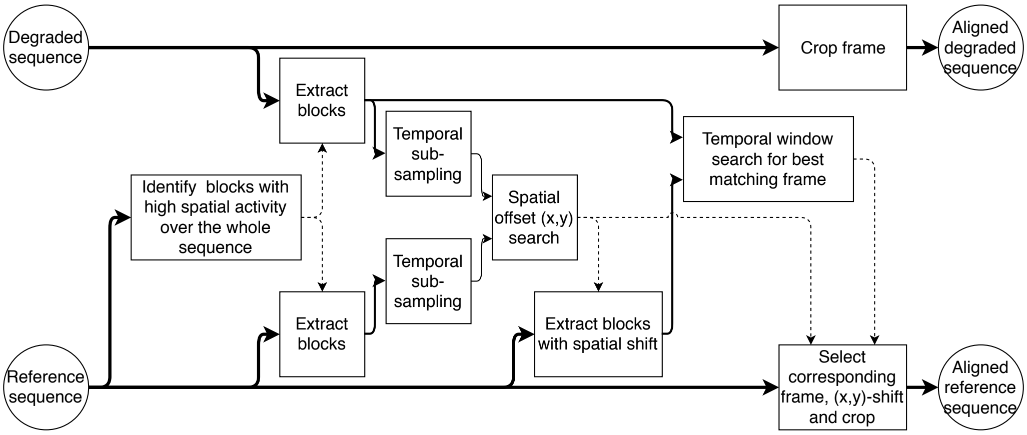
Degraded sequence
Crop frame
Aligned degraded sequence
Extract blocks
Temporal sub-sampling
Temporal window search for best matching frame
Identify blocks with high spatial activity over the whole sequence
Spatial offset (x,y) search
Temporal sub-sampling
Extract blocks
Extract blocks with spatial shift
Reference sequence
Select corresponding frame, (x,y)-shift and crop
Aligned reference sequence

Figure 2: Flow chart of the temporal registration process.

operates by computing and making available such measures for each frame in the video sequence. In addition, it also produces a single value expected to represent the quality of the whole sequence.

The usefulness of having different objective quality measures also stems from the fact that simple formulations such as the PSNR one, being a pure signal processing measure, disregards important factors such as viewing conditions and the human visual system (HVS) characteristics [28].

To overcome such limitations, other measures try to incorporate some natural visual characteristics. For instance, SSIM and its variants [24, 25] attempt to measure the structural distortion introduced in the distorted frame when compared to the reference one by combining a set of statistical measures such as mean, variance and covariance of the samples. The Visual Information Fidelity (VIF) [26] approach, instead, also relies on the use of Human Visual System (HVS) modeling as well as visual statistics, attempting to quantify the information that could be extracted by the observer from the reference image and the distorted one.

3.3. Temporal Pooling

Many video quality evaluation algorithms produce a single quality measure for each frame. Most algorithms also employ a simple temporal pooling strategy to obtain video quality measure of the whole sequence, i.e., the average the quality value of all frames in the sequence.

Such an approach presents the advantage of simplicity. However, parts of the video sequence presenting different quality levels should probably be weighted differently in the sequence quality measure. For instance, empirical observations showed that subjective scores are more influenced by the low quality parts than the high quality ones. Since there is currently no agreement in the research community about how to obtain a single quality measure from the values of each single frame, in our work we decided to experiment with several temporal pooling methods. Among them, we used some of the temporal pooling strategies that have been recently made available as part of the VMAF software distribution package [29], as well as some of the so-called "temporal collapsing functions" in the NTIA report [30] about video quality measurement techniques.

In particular, apart from the widely used *arithmetic* average, we computed the *median*, the *geometric* and *harmonic* averages, three *L^p-norm* (with $p = 1, 2, 3$), and the *75-th* and *90-th percentile* values. Assuming N measures x_i corresponding to the frames of the video sequence, we computed:

$$\begin{aligned} \text{arithmetic} &= \textstyle\sum_{i=1}^{N} \frac{1}{N} x_i \\ \text{geometric} &= \left(\textstyle\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}} \\ \text{harmonic} &= \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \\ \text{L}^p\text{-norm} &= \left(\textstyle\sum_{i=1}^{N} |x_i|^p\right)^{\frac{1}{p}} \end{aligned} \tag{1}$$

whereas the median and percentile first requires to sort the x_i values, then to select the one with the index i corresponding to $N/2$ (for the median) or the desired percentile P which corresponds to index $i = \frac{P}{100} N$. If a non-integer value is selected, the average values at the two nearest indexes is used.

4. Proposed Approach

4.1. Subset Selection

This section focuses on using different machine learning algorithms to select a set of representative PVSs from a larger one. Note that, in the literature, several machine learning algorithms have been used in the context of visual quality assessment. For instance, in [31, 32], Kmeans is used to cluster the large-scale database [33] to help select a representative subsets, i.e. small-scale subsets. In this work, instead, machine learning algorithms are used to find similarities or redundancies in a database of video sequences. The underlying principle is as follows. When machine learning algorithms assign several PVS to the same cluster, it is expected that they contain too many similarities to be challenging enough for the objective metrics; otherwise, the PVS is hard to assign and it can be quite challenging for the objective metrics.

As mentioned in Section 3, databases are subjectively evaluated and objectively evaluated with different pooling strategies as well. For each pooling strategy, we prepared a data matrix $X(m, n)$, where m and n respectively represents the number of samples, i.e., the PVSs in our case, and the number of features, i.e., the objective mapped scores in our case. $X(m, n)$ is prepared as illustrated in the blue part of Figure 3:

- The PVSs are temporally registered with their reference on a frame-by-frame basis.
- The FR objective scores are measured per frame, in particular PSNR, SSIM, MS-SSIM, and VIF.
- Different pooling strategies are applied to compute a single score for the whole video sequence for each metric.
- For the purpose of normalization, the objective scores are mapped to MOS scores (from 0 to 5) using a generalized logistic function with four fitted shape parameters.

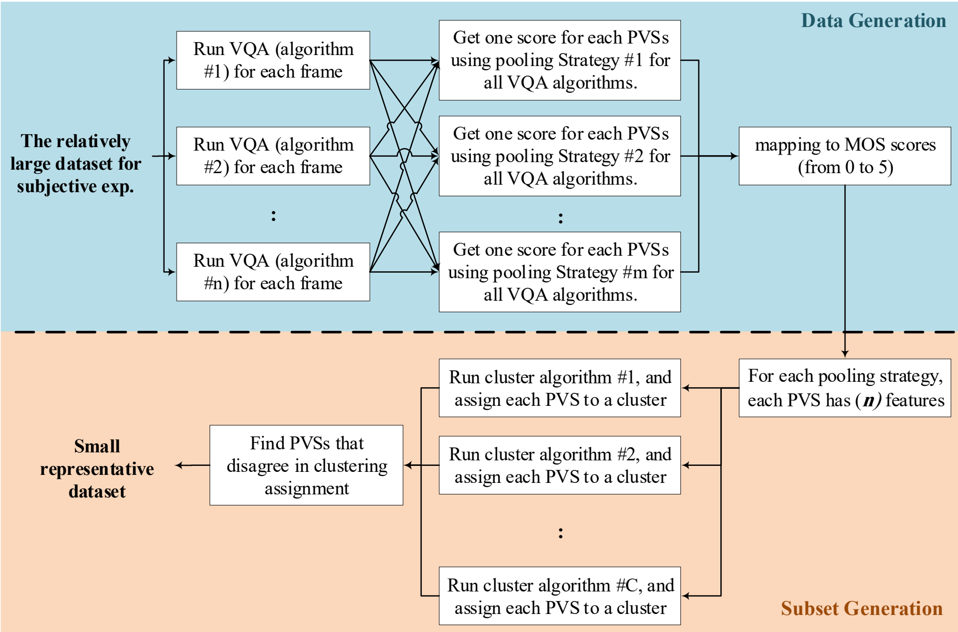
Data Generation
The relatively large dataset for subjective exp.
Run VQA (algorithm #1) for each frame
Run VQA (algorithm #2) for each frame
:
Run VQA (algorithm #n) for each frame
Get one score for each PVSs using pooling Strategy #1 for all VQA algorithms.
Get one score for each PVSs using pooling Strategy #2 for all VQA algorithms.
:
Get one score for each PVSs using pooling Strategy #m for all VQA algorithms.
mapping to MOS scores (from 0 to 5)
For each pooling strategy, each PVS has (*n*) features
Run cluster algorithm #1, and assign each PVS to a cluster
Run cluster algorithm #2, and assign each PVS to a cluster
:
Run cluster algorithm #C, and assign each PVS to a cluster
Find PVSs that disagree in clustering assignment
Small representative dataset
Subset Generation

Figure 3: The flow of preparing the data $X(m, n)$ and the generation of the small representative dataset (proposed methodology)

For instance, $X(m, n)$ for the mean pooling strategy contains *four* features. The first one is the mapped $PSNR\_mean$ score. The second, third, and fourth features are $SSIM\_mean$, $MS\text{-}SSIM\_mean$, and VIF_mean, respectively.

This data matrix $X(m, n)$ is fed into three unsupervised machine learning algorithms: Kmeans++ [34], Hierarchical clustering [35] and Gaussian mixture model (GMM). Kmeans is a clustering algorithms that aims to minimize the sum of point to centroid distances, Euclidean distance in our case. Hierarchical clustering aims to build a cluster tree following two main steps: the first is to measure the similarity between the observations, in our case through the Euclidean distance, and then the data are grouped using a linkage function, here the Ward's method. Finally, the Gaussian mixture model aims at maximizing the component of the posterior probability given the data. It can be considered as a soft Kmeans clustering.

However, for each machine learning algorithm some parameters have to be decided. For instance, the number of clusters in the case of Kmeans, the distance and linkage metrics for hierarchical clustering. In this paper, the data is clustered into $(K =)$ 3 clusters. Moreover, the Euclidean distance and the Ward's linkage metric are used in the hierarchical clustering.

After this decision, the selection process of the PVSs is rather straightforward. Once the clustering is done for each pooling strategy, the PVSs that are assigned to different clusters are counted to be the candidate PVSs and to be a representative subset. The red part of Figure 3 illustrates the process. Note that the size of the representative dataset depends on the clustering algorithm results, i.e. on comparing the similarity in clustering assignments, not on the clustering parameters such as number of clusters.

Note that the selection of the number of clusters is guided by the characteristics of the feature spaces of different objective measures. Classifying the data into two clusters is analogous to low and high quality groups, while classifying them into three clusters is analogous to low, medium, and high qualities. Hence, selecting more than three clusters is not reasonable due to the limitation of the feature space of the objective measures. In this paper, the data is clustered into two and three clusters.

4.2. Subset Comparison Methodology

To demonstrate the effectiveness of the proposed subset selection procedure, the proposed technique is compared to an extensive amount of random subsets. To make the comparison between different subsets, the analysis techniques as described in the VQEG-HDTV test report [18] have been applied.

Initially, different subset selection processes are performed to allow a proper comparison. From the progressive datasets, namely 423 PVSs from VQEG-HDTV datasets 1,3, and 5, the different subsets are selected based on the proposed solution. Additionally, in order to make a comparison, different subsets are generated in a random fashion. Random subsets of PVSs are selected ranging from a 5% reduction in dataset size (approx. 400 PVSs) up to 95% reduction (approx. 22 PVSs) in steps of 5%. From these 19 differently sized subsets, 100 random versions have been generated. As a consequence, the proposed subset selection will be compared with 1,900 random subset selection strategies.

If a 95% reduction in dataset size corresponds to only subjectively evaluating approx. 22 PVSs, it is clear that different conclusions will be drawn from such ~~ana~~ subjective test. To quantify this difference the following process is followed. The MOS from the different PVSs is transformed to a difference mean opinion score (DMOS) calculated as follows:

$$DMOS = MOS(PVS) - MOS(SRC) + 5 \tag{2}$$

In contrast to subjective scores, scores predicted by an objective model are not clipped at a minimum and maximum MOS. Therefore, it is crucial to map the objective score or VQR with a trend corresponding to the behavior of subjective scores. So, the VQRs are mapped using the following cubic polynomial function [36, 37]:

$$DMOS_p = a \cdot VQR^3 + b \cdot VQR^2 + c \cdot VQR + d \tag{3}$$

In this equation, $DMOS_p$ is the mapped/predicted DMOS and a, b, c, and d are calculated using a nonlinear programming solver optimizing for the least square error between the $DMOS_p$ and the actual DMOS [38]. As additional constraints, the optimization should lead to a monotonous increasing function with the inflection point outside the interval.

When the objective ratings are on the same scale as the DMOS values, the RMSE evaluation metric is calculated between the $DMOS_p$ of the objective metric and the actual DMOS. So, for every objective quality metric the following RMSE value is calculated:

$$\text{RMSE} \quad = \quad \sqrt{\tfrac{1}{N-d}\sum_{i=1}^{N}(DMOS(i) - DMOS_p(i))^2} \tag{4}$$

In this equation, N is the number of PVSs and d is the degrees of freedom of the mapping function, i.e. 4.

In VQEG's validation experiment, the objective measures were compared to each other in terms of RMSE values taking into consideration whether or not a statistical significant difference existed between any two RMSE values of the six tested algorithms. To evaluate different subsets, the significance of the differences in RMSE needs to be known and compared to the results found during the standardization process. It will be assumed that a subset is good if it leads to the same results concerning ranking and statistical difference.

With the assumption that the two populations are normally distributed, we define the H_0 hypothesis as there being no difference between the RMSE scores of two objective measures. To evaluate this hypothesis the following statistic is evaluated against the F-distribution:

$$\zeta \quad = \quad (\tfrac{RMSE_{max}}{RMSE_{min}})^2 \tag{5}$$

From the pair of RMSE values, $RMSE_{max}$ is the highest one and $RMSE_{min}$ the lowest one. When this ζ statistic is higher than F(0.05, N-d, N-d) within the 95% significance level, then there is a significant difference between the two RMSE scores.

This significance is calculated between the six objective measures evaluated in the VQEG-HDTV study [18] resulting in 15 comparisons. The result of calculating these 15 numbers on the entire dataset will be called the ground truth analysis result. When taking a subset of the dataset, the significance between different objective measures will change and these differences are summed. From here on, this number of differences with the full set will be called the Subset Error (SError). Only when this number is zero, the analysis of the subset will lead to the same conclusion as the analysis of the full dataset. The higher the SError, the more a subset will result in different conclusions about the comparison of the performance of different objective measures.

Finally, objective quality metrics can be pairwise ranked using the previously calculated significance information. A ranking between a pair of metrics is only meaningful when there is a significant difference between the two metrics. Only when a significant ranking turns into a significant but opposite ranking, it will be counted as a ranking error. Other changes, for example when a significant ranked pair becomes insignificant (regardless of the order of the two objective metrics), the error is solely counted as a SError.

5. Results

5.1. Subset Selection

In this section, the analysis results will be studied. Figure 4 and Figure 5 show the clustering of the three machine learning algorithms for arithmetic and harmonic temporal pooling strategies respectively. The clusters are visualized using two features (objective scores): PSNR and VIF. In both figures, it can be observed, in the Kmeans algorithm, the effect of the Euclidean distance between the points and the centroid. The farthest points (from the centroid) are assigned to different clusters using the hierarchical clustering: this is due the effect of the linkage functions. Comparing these two algorithms led to some differences in assigning points to clusters. On the other hand, in the GMM, it should be noted that there is interference between the clusters. Since the GMM compute the posterior probability, each point has a probability for each cluster (soft clustering) and it is assigned to the cluster with the highest probability (hard clustering). Hence, the interfered points have weights to be assigned to any clusters which make them very challenging for other clustering algorithms and that is why they are selected as candidate points for subset generations.

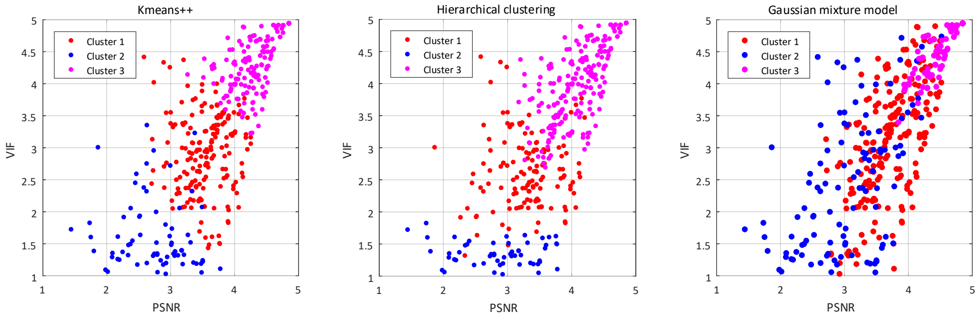

Figure 4: Clustering harmonic pooling strategy data using Kmeans, hierarchical clustering, and Gaussian mixture model with three clusters and with PSNR and VIF attributes.

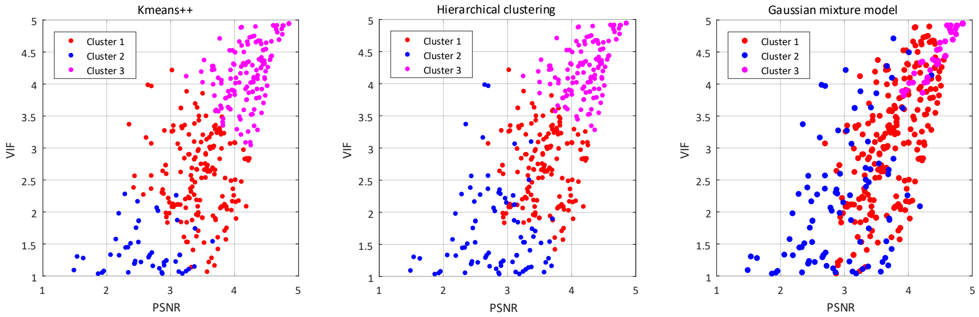

Figure 5: Clustering harmonic pooling strategy data using Kmeans, hierarchical clustering, and Gaussian mixture model with three clusters and with PSNR and VIF attributes.

Figure 6 and Figure 7 show the results of the proposed strategy for the arithmetic and the harmonic average pooling strategies, respectively. Each figure contains a plotting of each pair of features. Black points refer to the PVSs that are selected to be a representative subset when $k = 3$. With the arithmetic pooling strategy 225 of 423 PVSs are selected, whereas with the harmonic pooling strategy 175 of 423 PVSs are selected. Table 1 shows the number of selected PVSs for different number of clusters ($K = 2, 3,$ and, 4). It can be observed, from Figure 6 and Figure 7, that the selected PVSs are mainly located between the cluster borders in the harmonic pooling strategy, while this does not happen in the arithmetic pooling strategy. This can be seen in all sub-figures except PSNR/SSIM and PSNR/MS-SSIM. This could be explained by noting that, in

Table 1: The size of the selected PVSs for each pooling strategy for different number of clusters, $k = 2, 3$, and 4.

K	2	3	4
Size (mean)	45	225	228
Size (Geometric)	68	225	131
Size (Harmonic)	49	175	53
Size (Median)	60	194	210
Size (L^1-norm)	45	225	215
Size (L^2-norm)	59	210	205
Size (L^3-norm)	61	188	261

general, PSNR and SSIM seems to be highly related compared to the relation between other metrics. Therefore, using only PSNR and SSIM attributes is probably not enough for a good clustering.

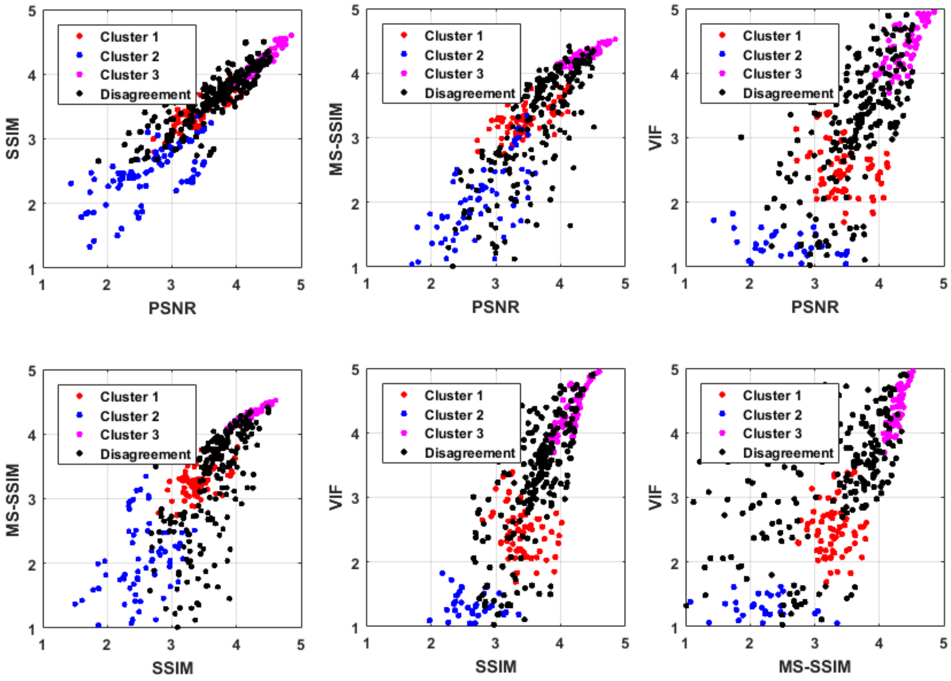

Figure 6: Clustering arithmetic pooling strategy data with three clusters. Black points are the candidate PVSs.

For the sake of completeness, it could be interesting to add the Mean opinion Score (MOS), in some form, to the input features, despite in a real-world scenario such value would not be known in advance. Therefore, for illustration purposes, a new data matrix $X(m, n)$ has been prepared with new attributes, i.e., the absolute differences between

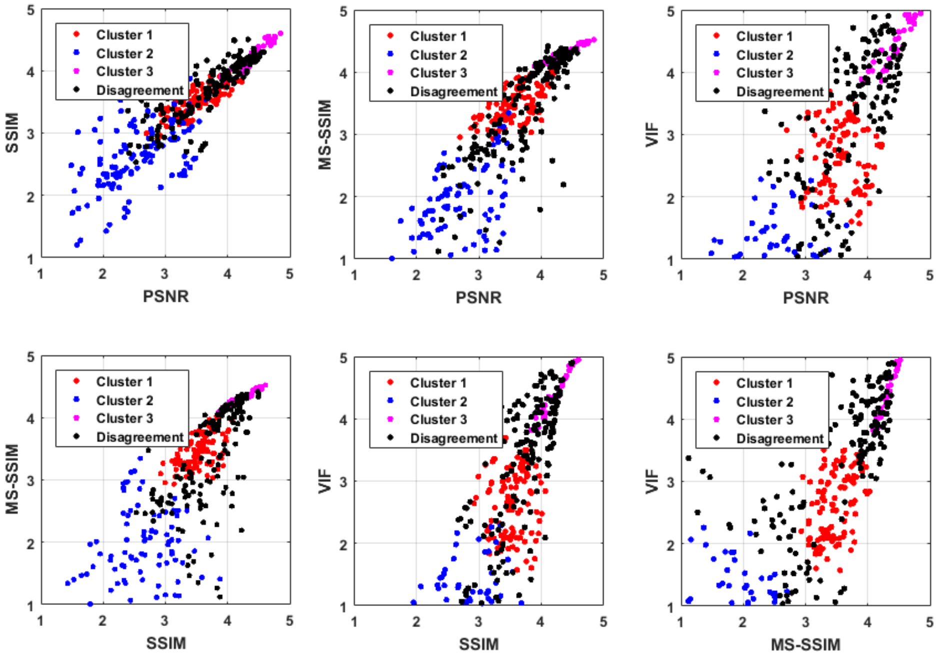

Figure 7: Clustering harmonic pooling strategy data with three clusters. Black points are the candidate PVSs.

the MOS scores and the objective scores. Figure 8 shows the results. It can be observed that the candidate PVSs (the black points) are located between the cluster borders. In addition, the figure also shows that the PSNR and SSIM are highly correlated compared to other metrics. Such behavior reinforces the previous observation for which using only PSNR and SSIM attributes is in general not enough for clustering.

5.2. Subset Comparison

In this section, the performance of an analysis of a subset that should give the same or similar results as the full dataset is investigated. The result of the analysis, described in Section 4.2, carried out on the full dataset can be found in Table 2. In this table, the RMSE results of the six different objective measures (A-F) can be found in the second row and column. There is also a significance label for each comparison between two objective measures. As an example, the table indicates that objective measure D performs significantly better than any other measure and measure C is significantly worse than any of the measures. When taking a subset results in a change of significance in this table, then such a smaller subjective evaluation would result in a different analysis result.

This table is also computed for the different subsets. As an example, the results of one of the best performing proposed subset selection procedures is provided in Table 3.

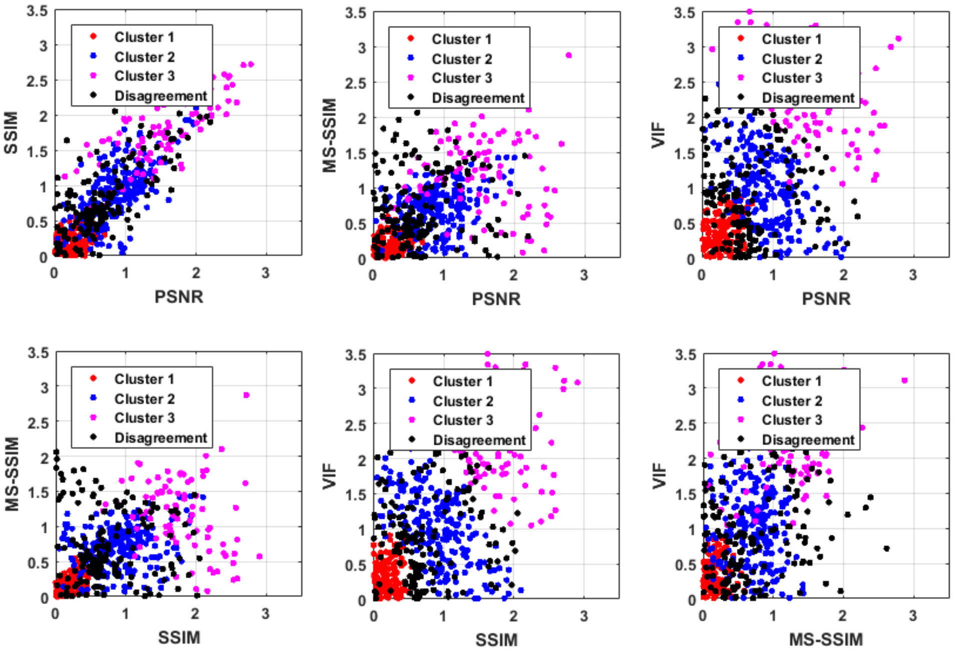

Figure 8: Clustering arithmetic pooling strategy data with three clusters and with new attributes; the absolute difference between the MOS scores and the objective scores. Black points are the candidate PVSs.

With respect to significance analysis, only a single significance change can be observed in the comparison between measures E and F ($SError = 1$).

Table 2: Ground truth analysis results as obtained from the full dataset. The RMSEs of six objective measures (A-F) are evaluated on the significance of their difference between each other.

		A	**B**	**C**	**D**	**E**	**F**
	RMSE	**0.71**	**0.78**	**0.99**	**0.57**	**0.72**	**0.79**
A	**0.71**		sign.	sign.	sign.	not	sign.
B	**0.78**			sign.	sign.	not	not
C	**0.99**				sign.	sign.	sign.
D	**0.57**					sign.	sign.
E	**0.72**						sign.
F	**0.79**						

To make a comparison between the proposed and random subsets, changes in significance are plotted in Figure 9. For the random subset selection, up to three SErrors can occur with as little as 5% reduction in dataset size. In the 75% to 85% dataset reduction region, between seven and ten of these analysis differences can be made in a worst case scenario. Although the best performing proposed subset selections results in one SError during the analysis, there is no random subset selection from the 100 tested possibilities

Table 3: Analysis results from the (raw OBJ) arithmetic dataset. Compared to the ground truth in Table. 2, the significance change between metrics E and F is in bold.

		A	**B**	**C**	**D**	**E**	**F**
	RMSE	**0.68**	**0.80**	**1.09**	**0.57**	**0.75**	**0.80**
A	**0.68**		sign.	sign.	sign.	not	sign.
B	**0.80**			sign.	sign.	not	not
C	**1.09**				sign.	sign.	sign.
D	**0.57**					sign.	sign.
E	**0.75**						**not**
F	**0.80**						

that performs better.

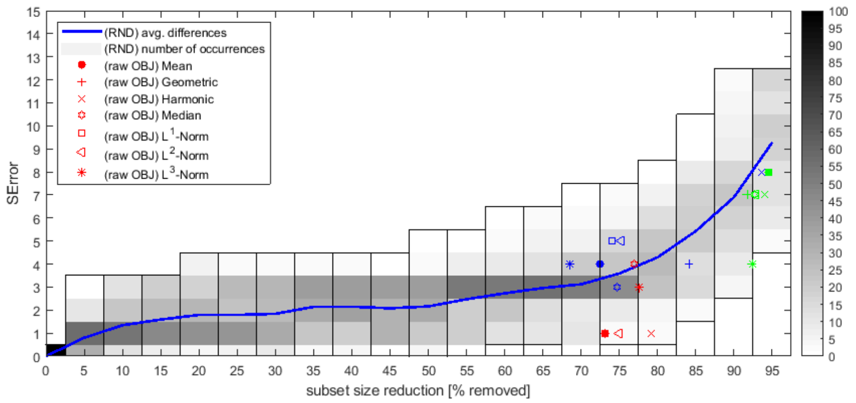

Figure 9: Amount of SErrors for 19 × 100 different random subsets (RND) selected from the full VQEG-HDTV dataset. Both average and color coded histogram of the SErrors are provided for the 100 random subsets of each subset size. The proposed subset selection is indicated as separate points in this plot for each pooling strategy and for different number of clusters, $k = 2$ (green), 3 (red), and 4 (blue).

From Figure 9, two main conclusions can be drawn. The first conclusion regards the performance of the generated/proposed subsets. All generated subsets with two and three clusters have a lower or equal SError compared to random subset selection. It is noticed that the datasets when having four clusters is not efficient because having more than three clusters is not reasonable due to the characteristics of the feature space. Hence, this effort could help in selecting the challenging PVSs for the subjective experiment. For instance, in this experiment, the PVSs are reduced to 175 out of 423 of the full subset when $k = 3$ and harmonic pooling strategy are used. The second conclusion regards the performance comparison of different pooling strategies. Comparing SErrors of subset selection, it can be noticed that some pooling strategies, in this case the arithmetic and geometric means, are performing better than others, notably the L^3-norm temporal pooling-based strategy.

Finally, analysis of the ranking behavior of the different subsets led to the following observations. All the proposed subset selection procedures result in the same ranking of the different objective measures as the full dataset analysis. The same observation can be made for the random subsets up to a subset reduction of 85%. Reducing the testset further leads to 2% of the random subsets to have one rank error. Reducing the subset size beyond 90% results in 10% of the subsets having a single rank error.

6. Discussion

Our approach has demonstrated that for the particular dataset VQEG-HDTV, similar results might have been obtained for the ITU Recommendation with only about one third of the effort of subjective assessment.

There are a few aspects that should be noted: First, our analysis is limited because it is a post-hoc analysis and its validity for planning a subjective validation experiment requires further work. The procedure would be to create a large set of PVS with many different SRC, preferably by performing online capture of the targeted video services (such as broadcast TV), then performing temporal alignment, running the simple objective measures, their temporal aggregations, and finally the subset analysis which would then potentially require a fusion of the generated subsets.

The second note concerns our evaluation procedure. While our comparison to random subset selection provides some hints that redundancies in the full set were identified, a more thorough analysis may be required, for example, by presenting the samples in each cluster to human observers who should identify similarities. In order to enable further research, we are providing our selected subsets online: (link will be inserted in the final version of the paper).

The third note is on the usage of VQEG-HDTV. We carefully selected this experiment for our evaluation because all PVS were created from SRC that were screened by all experts participating in the validation experiment and after the experiment, the PVS were again verified and validated to conform to the testplan. We are not aware of any other publicly available dataset that was screened so carefully. Nonetheless, it would be interesting to perform a similar analysis on other datasets.

Our approach should be seen as a generic procedure and it can be applied to other use cases. For example, in the evaluation of new coding algorithms, PVS may be selected in the same way. However, the set of similarity indicator (in our case, the simple quality measurement algorithms) may be amended by indicators that characterize the SRC with respect to coding parameters such as bitrate, length of motion vectors and other indicators that are typically used for bitstream analysis. Different indicators such as the Bjøntegaard's BD-PSNR [39] may also be included in the set.

7. Conclusions

In this work we tackled the problem of how to identify similarities in a video quality assessment dataset by means of objective quality measures processed through machine-learning based algorithms. Such an approach can be employed in many contexts that would traditionally require large-scale subjective experiments such as validation experiments by ITU or MPEG. Taking advantage of one of the largest, publicly available, subjectively annotated video quality dataset, i.e., VQEG-HDTV, we showed that the proposed approach yields to interesting reductions of the amount of work to be done in an actual subjective experiment without substantially modifying the conclusion that would be drawn from the larger, more expensive, experiment. In this experiment, the

PVSs are reduced to small size relative to the full subset which correspondingly reduce full experiment's time, money, and efforts.

References

[1] ITU-T Rec. J.144, Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference (Mar. 2004).

[2] ITU-T Rec. J.343, Hybrid perceptual bitstream models for objective video quality measurements (Nov. 2014).

[3] Video Quality Experts Group (VQEG), http://www.vqeg.org (May 2018).

[4] C. Keimel, T. Oelbaum, K. Diepold, Improving the verification process of video quality metrics, in: International Workshop on Quality of Multimedia Experience (QoMEX), San Diego, California, U.S.A., 2009, pp. 121–126.

[5] F. M. Ciaramello, A. R. Reibman, Systematic stress testing of image quality estimators, in: Image Processing (ICIP), 2011 18th IEEE International Conference on, 2011, pp. 3101–3104.

[6] A. R. Reibman, A strategy to jointly test image quality estimators subjectively, in: Image Processing (ICIP), 2012 19th IEEE International Conference on, 2012, pp. 1501–1504.

[7] U. Reiter, J. Korhonen, Comparing apples and oranges: Subjective quality assessment of streamed video with different types of distortions, in: International Workshop on Quality of Multimedia Experience (QoMEX), San Diego, California, U.S.A., 2009, pp. 127–132.

[8] S. Voran, A. Catellier, Gradient ascent paired-comparison subjective quality testing, in: International Workshop on Quality of Multimedia Experience (QoMEX), San Diego, California, U.S.A., 2009, pp. 133–138.

[9] M. Pinson, S. Wolf, An objective method for combining multiple subjective data sets, in: SPIE Video Communications and Image Processing Conference, 2003, pp. 8–11.

[10] M. Pinson, M. Barkowsky, P. Le Callet, Selecting scenes for 2D and 3D subjective video quality tests, EURASIP Journal on Image and Video Processing 2013 (1) (2013) 1.

[11] ITU-T Rec. P.910, Subjective video quality assessment methods for multimedia applications (Oct. 1996).

[12] A. Aldahdooh, E. Masala, O. Janssens, G. Van Wallendael, M. Barkowsky, Comparing simple video quality measures for loss-impaired video sequences on a large-scale database, in: Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on, IEEE, 2016, pp. 1–6.

[13] M. H. Pinson, S. Wolf, Techniques for evaluating objective video quality models using overlapping subjective data sets, NTIA Report TR-09-457.

[14] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, P. Le Callet, Aligning subjective tests using a low cost common set, in: QoE for Multimedia Content Sharing, Lisbon, Portugal, 2011.

[15] Y. Pitrey, R. Pépion, P. Le Callet, M. Barkowsky, Using overlapping subjective datasets to assess the performance of objective quality metrics on scalable video coding and error concealment, in: Proceedings of 2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX), Melbourne, Australia, 2012, p. 103.

[16] S. Zielinski, On some biases encountered in modern audio quality listening tests (part 2): Selected graphical examples and discussion, J. Audio Eng. Soc 64 (1/2) (2016) 55–74.

[17] S. Zielinski, F. Rumsey, S. Bech, On some biases encountered in modern audio quality listening tests — a review, Journal of the AES 56 (6) (2008) 427–451.

[18] Video Quality Experts Group, Report on the validation of video quality models for high definition video content (v. 2.0) (Jun. 2010).

[19] ITU-T Rec. J.341, Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference (Jan. 2011).

[20] ITU-T Rec. J.342, Objective multimedia video quality measurement of HDTV for digital cable television in the presence of a reduced reference signal (Apr. 2011).

[21] CDVL Technical Committee, The consumer digital video library, http://www.cdvl.org (May 2012).

[22] G. Cermak, L. Thorpe, M. Pinson, Test plan for evaluation of video quality models for use with high definition TV content, Video Quality Experts Group (VQEG).

[23] M. Barkowsky, J. Bialkowski, R. Bitto, A. Kaup, Temporal registration using 3D phase correlation and a maximum likelihood approach in the perceptual evaluation of video quality, in: IEEE 9th Workshop on Multimedia Signal Processing, IEEE, 2007, pp. 195–198. `doi:10.1109/MMSP.2007.4412851`.

[24] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.

[25] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, Vol. 2, 2003, pp. 1398–1402. `doi:10.1109/ACSSC.2003.1292216`.

[26] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2) (2006) 430–444.

[27] P. Hanhart, R. Hahling, Video quality measurement tool (VQMT), http://mmspg.epfl.ch/vqmt (Sep. 2013).

[28] L. Guo, Y. Meng, What is wrong and right with MSE?, in: Proc. 8th Int. Conf. Signal Image Process, 2006, pp. 212–215.

[29] Netflix, VMAF - video multi-method assessment fusion v.0.6.3, https://github.com/Netflix/vmaf (May 2018).

[30] S. Wolf, M. H. Pinson, Video quality measurement techniques, NTIA Report TR-09-457.

[31] A. Aldahdooh, E. Masala, G. Van Wallendael, M. Barkowsky, Framework for reproducible objective video quality research with case study on PSNR implementations, Digital Signal Processing 77 (2018) 195–206.

[32] A. Aldahdooh, E. Masala, G. Van Wallendael, M. Barkowsky, Reproducible research framework for objective video quality measures using a large-scale database approach, SoftwareX `doi:10.1016/j.softx.2017.09.004`.

[33] G. Van Wallendael, N. Staelens, E. Masala, M. Barkowsky, Full-HD HEVC-encoded video quality assessment database, in: Ninth International Workshop on Video Processing and Quality Metrics (VPQM), 2015.

[34] D. Arthur, S. Vassilvitskii, K-means++: The advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[35] J. H. Ward Jr, Hierarchical grouping to optimize an objective function, Journal of the American statistical association 58 (301) (1963) 236–244.

[36] ITU-T Rec. P.1401, Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models (Appendix II) (Jul. 2012).

[37] ITU-T Rec. J.149, Method for specifying accuracy and cross-calibration of video quality metrics (VQM) (Mar. 2004).

[38] C. Fenimore, J. Libert, M. Brill, Monotonic cubic regression using standard software for constrained optimization (Nov. 1999).

[39] G. Bjøntegaard, Improvements of the BD-PSNR model, ITU-T SG.16 Q.6 Doc. VCEG-AI11 (2008).