How to Train No Reference Video Quality Measures for New Coding Standards using Existing Annotated Datasets?

How to Train No Reference Video Quality Measures for New Coding Standards using Existing Annotated Datasets? *

Lohic Fotio Tiotsop
Dept of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
lohic.fotiotiotsop@polito.it

Tomas Mizdos
Dept of Multimedia and ICT
University of Zilina
Zilina, Slovakia
tomas.mizdos@feit.uniza.sk

Enrico Masala
Dept of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
enrico.masala@polito.it

Marcus Barkowsky
Deggendorf Institute of Technology
University of Applied Sciences
Deggendorf, Germany
marcus.barkowsky@th-deg.de

Peter Pocta
Dept of Multimedia and ICT
University of Zilina
Zilina, Slovakia
peter.pocta@feit.uniza.sk

***Abstract*—Subjective experiments are important for developing objective Video Quality Measures (VQMs). However, they are time-consuming and resource-demanding. In this context, being able to reuse existing subjective data on previous video coding standards to train models capable of predicting the perceptual quality of video content processed with newer codecs acquires significant importance. This paper investigates the possibility of generating an HEVC encoded Processed Video Sequence (PVS) in such a way that its perceptual quality is as similar as possible to that of an AVC encoded PVS whose quality has already been assessed by human subjects. In this way, the perceptual quality of the newly generated HEVC encoded PVS may be annotated approximately with the Mean Opinion Score (MOS) of the related AVC encoded PVS. To show the effectiveness of our approach, we compared the performance of a simple and low complexity but yet effective no reference hybrid model trained on the data generated with our approach with the same model trained on data collected in the context of a pristine subjective experiment. In addition, we merged seven subjective experiments such that they can be used as one aligned dataset containing either original HEVC bitstreams or the newly generated data explained in our proposed approach. The merging process accounts for the differences in terms of quality scale, chosen assessment method and context influence factors. This yields a large annotated dataset of HEVC sequences that is made publicly available for the design and training of no reference hybrid VQMs for HEVC encoded content.**

***Index Terms*—subjective experiment, data reuse, video quality, HEVC encoded videos, hybrid model, machine learning**

I. Introduction

New video codecs are being constantly proposed. In order to develop objective video quality measures (VQMs) capable of predicting the perceptual quality of video sequences processed with these new codecs, researchers typically need new subjectively annotated datasets. Unfortunately, conducting a subjective experiment might involve a significant cost in terms of time as well as financial resources. For this reason, this work focuses on approaches that allow to reuse the already existing subjectively annotated datasets for new coding standards.

The contribution of this paper is threefold and can be summarized as follows:

1) An approach to generate an annotated dataset of HEVC encoded video content from the existing data gathered during subjective experiments on AVC encoded PVSs is proposed.
2) A large-size annotated dataset of HEVC encoded sequences is built by merging the generated data and the data gathered in seven subjective experiment using the well-known Iterative Nested Least Square Algorithm (INLSA) [1]. The dataset is made freely available to researchers at http://media.polito.it/mmsp2021.
3) A low complexity no reference hybrid VQM is trained on the proposed dataset to validate its effectiveness. This VQM is meant to be used as a baseline to compare with when using the large HEVC dataset for designing hybrid models. Both its scores and the code for running the trained VQM are available at http://media.polito.it/mmsp2021.

In more details, our approach consists of generating HEVC encoded bitstreams and decoded PVSs with an attempt to make their perceptual qualities as similar as possible to those of given AVC encoded PVSs whose perceptual qualities have already been assessed during a subjective test. We have considered three subjective datasets containing AVC encoded

This work was partially supported by the PoliTO Interdepartmental Center for Service Robotics (PIC4Ser, http://pic4ser.polito.it); and by HPC@POLITO (http://www.hpc.polito.it).

978-1-6654-3288-7/21/$31.00

sequences. For each of these datasets, the subjective quality of the PVSs was assessed using different methods and scales. After calculating the value of three objective metrics, i.e. PSNR, SSIM and VIF, we implemented the INLSA to align the three datasets on the same quality scale, i.e. the Absolute Category Rating (ACR) scale ranging from 1 to 5. The alignment made by the INLSA takes into account the context influences factors of each experiment [1]. We then proposed an approach that exploits objective measures to estimate with which quantization parameter a source (SRC) content should be HEVC encoded in order to obtain a PVS that has the same perceptual quality as an AVC encoded PVS derived from the same SRC. Proceeding in this way, we converted the three original annotated datasets containing AVC encoded PVSs into a single one newly annotated dataset containing HEVC encoded PVSs. This dataset is referred to as "generated dataset" in the rest of this work.

We then considered the data from seven other subjective experiments but this time conducted directly on HEVC encoded PVSs. Although this data was already publicly available, they could not easily be jointly used for the training of VQMs for the quality assessment of HEVC encoded PVSs. This is because the subjective data of each dataset were collected in different contexts, while deploying different methods and reported on different scales. With these seven datasets we also created a single INLSA aligned dataset, which we will call "real dataset" in the paper, which takes into account the context influence factors of each experiment and reports all the subjective annotations on the ACR scale. The "generated dataset" and the "real dataset" were then merged together to propose a large annotated dataset of HEVC encoded PVSs, which we call the "GR-HEVC datatset", i.e. Generated and Real annotated HEVC data. It can be readily used by researchers to train, validate and evaluate the performance of VQMs.

To assess our proposal's effectiveness, we show that when the "generated dataset" derived from our approach is used as a training set, it is possible to train a model for the prediction of the perceptual quality of HEVC encoded PVSs that has an accuracy similar to that which would be obtained if the model would be instead trained on the "real dataset". To train the model, we extracted three simple and low complexity features from each PVS. The first two are widely used in no-reference models: the Quantization Parameter (QP) and Coding Unit size (CU), which are obtained from the bitstream information [2], while the last one, named Residual Energy (RE), is computed from the video pixel values. The RE is used in this work as a feature that accounts for the video temporal complexity. More details are provided in Section IV. The three features were then regressed to the quality scale using a shallow neural network. In many different testing conditions the model trained on the "generated dataset" shows performance comparable with those trained on the "real dataset". Finally, we observe that by jointly using the "generated dataset" and the "real dataset", as well as the three simple and low complexity features mentioned above, i.e. the QP, CU and RE, a no reference hybrid model can be derived. Its performance outperforms that of two well-known and widely used full reference metrics, i.e. SSIM and VIF in many testing conditions.

The remainder of the paper is organized as follows. In Section II the related work is briefly reviewed. Section III provides a detailed presentation of the proposed approach to generate data for new coding standards by deploying existing annotated datasets. In Section IV, a description of the feature extraction process for training the simple model used to validate our approach is presented. Section V is devoted to the results presentation and discussion, finally conclusions are drawn in Section VI.

II. Related work

Several objective VQMs have been proposed in the last decades [3]. Along with these VQMs, many subjectively annotated datasets have been published. However, two major obstacles hinder the effective use of these datasets. On the one hand, due to the fact that subjective experiments are time consuming and resources demanding, individual subjectively annotated datasets publicly available are generally not large in terms of size. On the other hand, since there is no consensus on the fact that a certain methodology for conducting subjective experiments is better than all the others, the existing datasets contain subjective scores in different formats. This makes the joint use of existing datasets for training supervised machine learning based VQMs difficult. Many papers have proposed solutions to address these two problems.

Regarding the first issue, i.e. the limitation in terms of size of subjectively annotated datasets, some authors used approaches that would allow them not to overfit the training set, e.g. transfer learning and shallow neural networks, to propose machine learning (ML) based VQMs [4] and other tools useful for modeling the diversity of end-users opinions [5]. The resulting quality estimators are mainly restricted only to the use-cases included in the dataset used for their training. Since the datasets are usually small in size, these metrics generally have limited application scopes. For this reason, other authors have instead chosen to enlarge the datasets by adding other PVSs whose quality is annotated using full reference VQMs. They thus obtain a training set in which one part of the PVSs has been subjectively evaluated and the subjective MOS is available, while the other part has been only objectively evaluated with full reference metrics and the related scores were mapped to the MOS scale. For instance, in [6] the authors used the score of the SSIMplus [7], as labels instead of the MOS to enlarge the training set. In many other works [8]–[10], the authors did not use any subjective annotations, they have proposed to predict the scores of some full reference VQMs, i.e. PSNR. SSIM, and VIF, by deploying no reference features. In this way, very large datasets can be created as the labels are provided by an algorithm, i.e. the chosen full reference VQMs.

Some approaches have also been proposed to tackle the second issue, i.e. how to deal with the differences in terms of

quality scale and context influence factors in order to jointly exploit existing subjectively annotated datasets. In particular, in [1] the authors proposed the INLSA. It is an algorithm that is meant to allow a fusion of many different subjectively annotated datasets to produce a single one in which the subjective scores are reported on the same quality scale. The INLSA algorithm was designed to account for the potential bias due to the way the participants used the quality scale in the individual experiments following the instructions provided in the training phase. Therefore, it accounts for the context influence factors of each experiment up to a certain extent. The authors in [11] proposed an approach to train a neural network based VQM by jointly using subjective data gathered in different contexts and with different quality scales. The authors of [12] proposed an approach to map the results of experiments conducted in the form of pair comparison to a continuous quality scale.

The papers mentioned so far propose approaches to effectively use existing subjectively annotated datasets containing PVSs generated by a certain video coding standard in order to develop VQMs capable of predicting the perceptual quality of video content processed by the same coding standard. This work, instead, aims to evaluate the possibility of using existing data on previous video coding standards to generate useful data for a training of VQMs able to face challenges introduced by newer codecs. We focus in particular on how to use existing subjectively annotated datasets involving AVC encoded content to generate useful data for a training of VQMs able to predict the perceptual quality of HEVC encoded video content. More precisely, given an AVC encoded PVS for which a subjective evaluation is available, we search for the optimal QP that should be used by an HEVC encoder to compress the source of that PVS in order to generate an HEVC encoded sequence that is expected to have the same perceptual quality as the AVC encoded PVS. In a recent journal paper [13], we have shown the effectiveness of this approach on still images.

After showing the effectiveness of the "generated dataset", we performed some analysis to merge it with seven subjectively annotated datasets involving HEVC encoded content. We showed that the obtained dataset is a suitable asset for the research community as a very simple no reference hybrid model trained on it outperformed full reference metrics in many test conditions. Hybrid models for HEVC encoded content are still of high interest [14].

III. Generating Annotated Datasets for New Video Coding Standards

This section presents our approach to generate an annotated dataset of HEVC encoded PVSs using existing subjective data of AVC encoded PVSs.

Given a H.264/AVC encoded PVS, we first computed its PSNR, SSIM and VIF quality predictions. Afterward, from the SRC of that PVS we generated 52 HEVC encoded sequences by means of the HM reference software [15] choosing the QP ranging from 0 to 51, then we computed also the PSNR, SSIM and VIF quality predictions for all those 52 newly created sequences. We identified, among the 52 sequences, the one whose PSNR value is closest to that of the H.264/AVC encoded PVS. This encoded sequence has a given QP that we denote as QP_PSNR. Analogously, we determined the values of QP_SSIM and QP_VIF. Then, we considered the median of these three QP values as an estimation of the QP to be used with the HEVC encoder in order to obtain a PVS whose perceptual quality is similar to that of the initial H.264/AVC encoded PVS. The quality attributed to the newly created HEVC encoded PVS is therefore assumed to be equal to the MOS of the related H.264/AVC encoded PVS. We experimentally found that the median of the three QP values was the best choice. In fact, by using it, the "generated dataset" allowed us to train a model whose accuracy is very close to that obtained by using data gathered in a subjective test. The Algorithm 1 summarizes the steps implemented by the proposed approach.

Algorithm 1: Data generation algorithm

Input: PVS_{AVC}, SRC;
$\text{diff}_{PSNR} = \infty$, $\text{diff}_{SSIM} = \infty$, $\text{diff}_{VIF} = \infty$;
$\text{PSNR}_{AVC} = \text{computePSNR}(\text{PVS}_{AVC}, \text{SRC})$;
$\text{SSIM}_{AVC} = \text{computeSSIM}(\text{PVS}_{AVC}, \text{SRC})$;
$\text{VIF}_{AVC} = \text{computeVIF}(\text{PVS}_{AVC}, \text{SRC})$;
for *QP = 0,1,...,51* **do**
  $\text{PVS}_{HEVC,QP} = \text{HEVC\_encode}(\text{SRC}, \text{QP})$;
  $\text{PSNR}_{QP} = \text{computePSNR}(\text{PVS}_{HEVC,QP}, \text{SRC})$;
  $\text{SSIM}_{QP} = \text{computeSSIM}(\text{PVS}_{HEVC,QP}, \text{SRC})$;
  $\text{VIF}_{QP} = \text{computeVIF}(\text{PVS}_{HEVC,QP}, \text{SRC})$;
  if $\mid PSNR_{AVC} - PSNR_{QP} \mid < diff_{PSNR}$ **then**
    $\text{QP}_{PSNR} = \text{QP}$;
    $\text{diff}_{PSNR} = \mid \text{PSNR}_{AVC} - \text{PSNR}_{QP} \mid$;
  end
  if $\mid SSIM_{AVC} - SSIM_{QP} \mid < diff_{SSIM}$ **then**
    $\text{QP}_{SSIM} = \text{QP}$;
    $\text{diff}_{SSIM} = \mid \text{SSIM}_{AVC} - \text{SSIM}_{QP} \mid$;
  end
  if $\mid VIF_{AVC} - VIF_{QP} \mid < diff_{VIF}$ **then**
    $\text{QP}_{VIF} = \text{QP}$;
    $\text{diff}_{VIF} = \mid \text{VIF}_{AVC} - \text{VIF}_{QP} \mid$;
  end
end
$\text{QP}_{HEVC} = \text{median}(\text{QP}_{PSNR}, \text{QP}_{SSIM}, \text{QP}_{VIF})$;
$\text{PVS}_{HEVC} = \text{HEVC\_encode}(\text{SRC}, \text{QP}_{HEVC})$;
Return: PVS_{HEVC}

The motivation for this algorithm is as follows. While subjective experiments remain the gold standard for obtaining annotated datasets with MOS, the proposed algorithm is expected to perform better than a simple annotation involving objective measurement algorithm. The performance of VQMs is known to depend strongly on the video content. It has been shown in various publications that the same coding algorithm, also called Hypothetical Reference Circuit (HRC), with different parameters, such as bitrate, leads to a high correlation between VQM and MOS when using the same SRC but to a low correlation when correlating across SRC [16], [17]. Our approach has the advantage that it is comparing

two HRCs, namely AVC and HEVC video coding, which cause perceptually similar artifacts, mostly blockiness and blurriness. The VQMs involved in our case are therefore used in their optimal application scenario, i.e. not comparing across SRC and judging the quality of similar HRCs. Although in this work we relied only on three well know VQMs with different characteristics, Algorithm 1 can be implemented by using more than three VQMs.

We have applied the Algorithm 1 to the H.264/AVC encoded PVSs used in three subjective experiments, i.e. VQEG HDTV [18], LIVE mobile [19] and IVP [20], to generate HEVC encoded PVSs whose perceptual quality is annotated with MOS values already available in the three aforementioned datasets. We then implemented an instance of the INLSA to align the MOS values accounting for the context influence factors of each experiments. This results in the "generated dataset" containing 118 HEVC encoded PVSs with their perceptual quality scores reported on the five point ACR scale.

We then combined the "generated dataset" with the results of seven subjective experiments on HEVC encoded video content in order to generate a large annotated dataset of HEVC encoded content. More precisely, we collected the results of seven subjective experiments, i.e. BVI-HD [21], BVI-VCE [22], BVI-Texture [23], SJTU-UHD [24], SJTU-FHD [24], UV5G [25] and BC [26], during which subjects were asked to evaluate the quality of HEVC encoded sequences. These experiments were not conducted using similar methods and quality scales. We also aligned the MOS values using the INLSA to create the "real dataset". Aligning the "real dataset" with the "generated dataset" yielded the GR-HEVC dataset that contains a total of 429 HEVC encoded PVSs associated with the aligned MOS values expressed on the ACR scale ranging from 1 to 5.

IV. Training a Simple Model for Verifying our Approach

As done in our previous work [13] on still images, the best way to assess the effectiveness of the "generated dataset" derived by our approach would be to show the generated HEVC encoded sequences to a set of subjects and see if the obtained MOS is the same as the MOS of the AVC encoded PVSs, which we used as quality labels. Unfortunately, this approach is time consuming and expensive. For this reason, as a preliminary solution to evaluate the effectiveness of the our approach, we show that the "generated dataset" deployed as a training set performs similar to the "real dataset" used in the same context. Therefore we trained a ML based no reference hybrid model, first using only the "generated dataset", and then the "real dataset". We then compared the performance of the obtained models.

For each PVS, we extracted three features. The QP and CU size were directly extracted from the bitstream information. For each frame of each PVS, the RE, which represents a pixel based feature, was computed as follows:

$$RE_f = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(P_f(i,j) - P_{f-1}(i,j)\right)^2 \qquad (1)$$

where I and J are the height and width of each frame, respectively. f is an index of the frame and $P_f(i,j)$ is a pixel value of the frame f positioned on row i and column j.

Note that the RE allows to capture the complexity of the PVS in terms of how fast the frame content is varying. Therefore, it considers up to certain extent the objects motion or scene changes that are well known to cause a temporal masking effect, which has a direct impact on the human's evaluation of the perceived visual quality. In most of the video encoder implementations, the CU size is strongly correlated to the average amount of spatial details within a given image area as larger CU sizes are used for flat regions for rate/distortion optimization. Taking into account the amount of spatial details while assessing the quality objectively yields predictions with a higher accuracy [4]. Finally, through the QP, we aimed at taking into account the impairment of the visual quality due to quantization artifacts. A higher QP usually results in blocking artifacts but, in the case of HEVC it may also lead to blurring artifacts because of a stronger impact of the in-loop filter. The QP thus focuses on modeling of the impact of the HRC, while the RE and CU focus on the SRC.

We computed the mentioned features for each frame of the examined PVS. To get only three parameters per PVS, we experimentally found that the best pooling strategies are as follows: for the QP and CU distribution, the average value over the frames, while for the RE, the 80% quantile of the sample of the frame values, i.e. $RE = quantile_{80\%}\{RE_f \quad f = 1, 2, \ldots, F\}$, where F is the total number of frames of the PVS. To obtain our validation model, we regressed the three aforementioned features to the MOS scale using ML based models.

Please note that our primary aim is not to train a VQM that is better than the state-of-the-art models but rather to demonstrate that both the "generated dataset" and the GR-HEVC dataset proposed in this paper represent valid annotated datasets for a training and validation of VQMs for HEVC encoded content. For this reason we used only the three simple features mentioned above.

V. Results

We have conducted numerical experiments to assess the effectiveness of the algorithm 1 and the GR-HEVC dataset. The results are discussed in more detail in upcoming subsections.

A. Effectiveness of the proposed data generation algorithm

To validate the effectiveness of Algorithm 1, we showed that the dataset it generated competes very well with the "real dataset" when used as training set. We have trained a no reference hybrid model by regressing the three features described in Section III to the quality scale with an artificial Neural Network (NN). The model was first trained using only the "generated dataset" as a training set and tested on the seven

TABLE I

COMPARING THE PERFORMANCE OF THE "GENERATED DATASET" TO THAT OF THE "REAL DATASET" AS TRAINING SETS. THE LETTER T STANDS FOR "TRAINING".

Dataset	T on "real dataset"			T on "generated dataset"		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
BVI-HD	0.86	0.85	0.40	0.86	0.85	0.41
BVI-Texture	0.85	0.85	0.27	0.78	0.76	0.32
BVI-VCE	0.65	0.60	0.29	0.86	0.88	0.20
SJTU-FHD	0.84	0.82	0.27	0.88	0.89	0.28
SJTU-UHD	0.85	0.81	0.18	0.83	0.78	0.26
UVG	0.72	0.51	0.27	0.64	0.40	0.29
BC	0.97	0.96	0.04	0.98	0.98	0.05

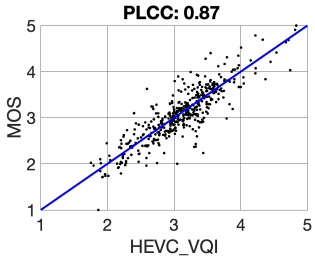

Fig. 1. Performance of the released HEVC_VQI on the GR-HEVC dataset

datasets that were merged to form the "real dataset". Then, seven similar models were trained using the "real dataset". To train each of these seven models, we used as a training set, six datasets out of the seven datasets that form the "real dataset". The resulting model was then tested on the dataset left out of the training process.

The results presented in Table I compare the performance of the model trained only on the "generated dataset" and that of the seven models trained on the "real dataset". As it can be noticed, the model trained with the data generated by algorithm 1, i.e. the "generated dataset", compares very well with the seven models trained on the data gathered during the pristine subjective tests. There are testing conditions for which the model that learned from the "generated dataset" outperformed the one trained on the "real dataset" and vice versa. This shows that the proposed approach is able to generate valid annotated datasets to be prospectively used for the training of VQMs for HEVC encoded content.

B. Effectiveness of the GR-HEVC dataset

We proved the effectiveness of the GR-HEVC dataset by showing that it can be used to train a low complexity no reference hybrid VQM that competes very well with two well-known and widely used full reference VQMs. We tested the performance of three different ML regression models. More precisely, we mapped the three features, i.e. the QP, the CU and the RE, to the MOS by training a regression tree (RT), support vector regression (SVR) model using the radial basis function as the kernel and finally shallow NN having a single hidden layer with three neurons. We call HEVC Video Quality Index (HEVC_VQI) any VQM obtained by regressing the three aforementioned features to the MOS scale. The performance of the three considered ML models in terms of Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC) and Root Mean Square Error (RMSE) can be seen respectively in Table II, III, IV. The models were trained using samples coming from nine datasets out of the 10 that form the GR-HEVC dataset, and tested on the one left out during the training process, i.e. the one reported in the table. In light of the results in Table II, III, IV, it is clear that the NN based model performed in general better than other ML regression methods. Therefore, the final no reference hybrid HEVC_VQI that we are releasing was obtained by training a NN using all the data available in the GR-HEVC dataset. Figure 1 shows the performance of the released VQM on the GR-HEVC dataset, i.e. its training set.

TABLE II

THE PROPOSED GR-HEVC DATASET ALLOWED TO TRAIN A NO REFERENCE METRIC *(HEVC_VQI)* THAT COMPARES WELL WITH TWO FULL REFERENCE ONES IN TERMS OF PLCC.

Dataset	Full ref VQMs		**HEVC_VQI**		
	SSIM	VIF	RT	SVR	NN
BC	0.95	0.93	0.18	**0.99**	0.97
BVI-HD	0.61	0.81	0.79	**0.87**	**0.87**
BVI-Texture	0.49	0.79	0.79	0.82	**0.86**
BVI-VCE	0.78	0.93	0.74	0.88	**0.94**
IVP	0.56	0.86	0.77	0.86	**0.87**
LIVE-mobile	0.59	0.91	0.65	0.89	**0.94**
SJTU-FHD	0.84	0.81	0.78	**0.87**	0.83
SJTU-UHD	0.76	0.81	0.76	0.80	**0.84**
UVG	0.65	0.69	0.56	**0.75**	0.72
VQEG-HDTV5	0.50	0.75	0.68	0.78	**0.84**

TABLE III

THE PROPOSED GR-HEVC DATASET ALLOWED TO TRAIN A NO REFERENCE METRIC *(HEVC_VQI)* THAT COMPARES WELL WITH TWO FULL REFERENCE ONES IN TERMS OF SROCC.

Dataset	Full ref VQMs		**HEVC_VQI**		
	SSIM	VIF	RT	SVR	NN
BC	0.95	0.81	0.31	**0.96**	**0.96**
BVI-HD	0.71	0.78	0.80	0.87	**0.88**
BVI-Texture	0.56	0.80	0.79	0.82	**0.85**
BVI-VCE	0.73	0.91	0.66	0.85	**0.94**
IVP	0.70	**0.88**	0.76	0.82	0.83
LIVE-mobile	0.73	0.92	0.61	0.90	**0.97**
SJTU-FHD	0.78	0.82	0.79	**0.83**	0.82
SJTU-UHD	0.69	0.78	0.79	0.81	**0.82**
UVG	0.51	0.50	0.49	**0.59**	0.57
VQEG-HDTV5	0.74	**0.81**	0.61	0.69	0.76

We compared the performance of the trained models, (one for each regression method) to that of two full reference VQMs, i.e. the SSIM and VIF. The results are summarized in Table II, III, IV that present, respectively, the PLCC, SROCC and RMSE of each VQM with respect to the MOS. Despite the fact that the proposed HEVC_VQI is a no reference metric, it shows quite competitive performance when compared to the full reference metrics as demonstrated by the results. For instance, the HEVC_VQI outperformed both full reference metrics in terms of the PLCC on all the 10 datasets that form the GR-HEVC dataset. We also note that the HEVC_VQI indeed yielded a significantly higher prediction accuracy than that of the SSIM for many testing conditions. These results highlights the suitability of the proposed GR-HEVC dataset as a valid set of data for prospective training of VQMs for HEVC encoded video content.

We would like to recommend that future research using the GR-HEVC dataset not only reports performance indicators such as PLCC, SROCC, and RMSE but also checks for statistical significance of improvements when comparing to

TABLE IV
THE PROPOSED GR-HEVC DATASET ALLOWED TO TRAIN A NO REFERENCE METRIC *(HEVC_VQI)* THAT COMPARES WELL WITH TWO FULL REFERENCE ONES IN TERMS OF RMSE.

Dataset	Full ref VQMs		**HEVC_VQI**		
	SSIM	VIF	RT	SVR	NN
BC	0.22	0.16	0.37	0.23	**0.06**
BVI-HD	0.72	0.47	0.58	0.58	**0.39**
BVI-Texture	0.48	0.33	0.35	0.30	**0.27**
BVI-VCE	0.39	0.15	0.36	0.19	**0.13**
IVP	0.33	0.19	0.35	0.24	**0.17**
LIVE-mobile	0.46	0.22	0.48	0.28	**0.14**
SJTU-FHD	0.37	0.36	0.39	**0.25**	0.34
SJTU-UHD	0.39	0.28	0.34	0.44	**0.21**
UVG	0.49	0.29	0.35	**0.27**	**0.27**
VQEG-HDTV5	0.42	0.20	0.28	0.20	**0.17**

the appropriate algorithm class, i.e. SSIM and VIF for Full Reference and HEVC_VQI for Hybrid Models.

VI. CONCLUSION

In this work we have proposed an approach to generate an annotated dataset of HEVC encoded video content deploying existing subjective data of AVC encoded PVSs. The proposed approach extends to the more generic case of using existing data on previous coding standards to address the challenges introduced by new codecs instead of designing and running new subjective experiments, which are both time consuming and resources demanding. We have shown that the data generated with our approach represents a training set that performs comparable to that containing subjective data obtained for the corresponding new video coding standard, i.e. HEVC. Moreover, we put together the generated data coming from three datasets representing the previous coding standard and the HEVC data collected during seven subjective experiments while accounting for the effect of each experiment's context influence factors to propose a larger dataset to be prospectively used for training of VQMs for HEVC encoded content. This dataset may be considered as a valuable asset for the research community, since the very simple no reference VQMs trained on it have a good performance compared to two full reference VQMs for many test conditions.

REFERENCES

[1] M. H. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," in *Visual Communications and Image Processing*, vol. 5150, 2003, pp. 583–592.

[2] T. Mizdos, M. Barkowsky, M. Uhrina, and P. Pocta, "Linking bitstream information to QoE: A study on still images using HEVC intra coding," *Advances in Electrical and Electronic Engineering*, vol. 17, no. 4, Dec. 2019.

[3] M. Vranješ, S. Rimac-Drlje, and K. Grgić, "Review of objective video quality metrics and performance comparison using different databases," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 1 – 19, 2013.

[4] L. Fotio Tiotsop, A. Servetti, and E. Masala, "Full reference video quality measures improvement using neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2737–2741.

[5] L. F. Tiotsop, T. Mizdos, M. Uhrina, M. Barkowsky, P. Pocta, and E. Masala, "Modeling and estimating the subjects' diversity of opinions in video quality assessment: a neural network based approach," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3469–3487, 2021.

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

[7] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," in *Human Vision and Electronic Imaging XX*, vol. 9394. International Society for Optics and Photonics, 2015, p. 939406.

[8] B. Lee and M. Kim, "No-reference psnr estimation for hevc encoded video," *IEEE Transactions on Broadcasting*, vol. 59, no. 1, pp. 20–27, 2013.

[9] M. A. Aabed and G. AlRegib, "No-reference quality assessment of hevc videos in loss-prone networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2015–2019.

[10] M. Shahid, J. Panasiuk, G. Van Wallendael, M. Barkowsky, and B. Lövström, "Predicting full-reference video quality measures using hevc bitstream-based no-reference features," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–2.

[11] L. Krasula, Y. Baveye, and P. Le Callet, "Training objective image and video quality estimators using multiple databases," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 961–969, 2020.

[12] M. Pérez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE Transactions on Image Processing*, vol. 29, pp. 1139–1151, 2020.

[13] T. Mizdos, M. Barkowsky, M. Uhrina, and P. Pocta, "How to reuse existing annotated image quality datasets to enlarge available training data with new distortion types," *Multimedia Tools and Applications*, pp. 1–23, 2021.

[14] A. Raake, S. Borer, S. M. Satti, J. Gustafsson, R. R. R. Rao, S. Medagli, P. List, S. Göring, D. Lindero, W. Robitza, G. Heikkilä, S. Broom, C. Schmidmer, B. Feiten, U. Wüstenhagen, T. Wittmann, M. Obermann, and R. Bitto, "Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204," *IEEE Access*, vol. 8, pp. 193 020–193 049, 2020.

[15] "HM reference software (v. 16.20)," 2015. [Online]. Available: https://hevc.hhi.fraunhofer.de/

[16] T. Oelbaum, K. Diepold, and W. Zia, "A generic method to increase the prediction accuracy of visual quality metrics," in *Picture Coding Symp.(PCS)*, 2007.

[17] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

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

[19] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[20] Image and Video Processing (IVP) Lab, The Chinese University of Hong Kong, "The IVP subjective quality video database," 2011. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/

[21] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2620–2630, 2018.

[22] A. V. Katsenou, F. Zhang, M. Afonso, and D. R. Bull, "A subjective comparison of AV1 and HEVC for adaptive video streaming," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4145–4149.

[23] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. Bull, "A video texture database for perceptual compression and quality assessment," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 2781–2785.

[24] J. Bienik, M. Uhrina, M. Vaculik, and T. Mižďoš, "Perceived quality of full HD video - subjective quality assessment," *Advances in Electrical and Electronic Engineering*, vol. 14, pp. 437–444, 2016.

[25] J. Bienik, M. Uhrina, and P. Kortis, "Influence of bit depth on subjective video quality assessment for high resolutions," *Advances in Electrical and Electronic Engineering*, vol. 15, no. 4 Special Issue, pp. 683–691, 2017.

[26] M. Uhrina, J. Bienik, M. Vaculik, and M. Voznak, "Subjective video quality assessment of VP9 compression standard for full HD resolution," in *2016 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, 2016, pp. 1–5.