Full Reference Video Quality Measures Improvement using Neural Networks

FULL REFERENCE VIDEO QUALITY MEASURES IMPROVEMENT USING NEURAL NETWORKS

Lohic Fotio Tiotsop *Antonio Servetti* *Enrico Masala*

Control and Computer Engineering Department
Politecnico di Torino, Torino, Italy

ABSTRACT

The accuracy of video quality metrics (VQMs) is an important issue for several applications. In this work, first we observe that the accuracy of several video quality metrics (VQMs) is strongly related to the spatial complexity index (SI) of the source. In particular, our investigation suggests that the VQMs are more likely to inaccurately predict the subjective quality of the processed video sequences derived from sources characterized by low SI. To address such a situation, we propose a machine learning based improvement for each of the VQMs considered in this work and a video quality metric fusion index (VQMFI) that jointly exploits all the VQMs considered in the study as well as spatiotemporal features to produce a better estimation of the subjective quality. Computational results demonstrate the superiority of our proposals on several datasets.

Index Terms— Video quality, machine learning, spatial activity

1. INTRODUCTION

Several objective video quality measures (VQMs) have been proposed in the past decades to compute an estimate of the mean opinion score (MOS) that would be produced in a subjective video quality experiment in which users are asked their opinion about the quality of videos affected by artifacts, the so called processed video sequences (PVSs), when compared to the original source (SRC).

Despite many of those measures have been shown to perform relatively well in a wide range of conditions, the correlation between the predicted MOS and the actual one is still far from optimal. In particular, in this work we start from the observation that sources with low spatial activity seem to negatively affect the correlation between the value of the measure and the actual MOS.

On the basis of such an observation we propose a correction to the behavior of such measures by means of a neural network (NN) based approach. Hence the goal of this work is not to design a highly sophisticated NN for quality assessment, as already done by several other authors, instead we aim at designing simple and very low-complexity NNs that allow to enhance the accuracy of the traditional VQMs. Such NNs take as input the traditional measure as well as two spatiotemporal features and returns an *improved* quality value that is shown to yield a better MOS estimation.

Moreover, we also use the same NN based approach to create a new *fused* measure that, relying on multiple well-known objective quality measures, produces a value that is shown to better correlate with the subjective video quality also when used on a completely different video quality dataset. The performance is even better than VMAF [1] (the best measures among the ones considered in this work) for the case of low spatial activity and almost on par in the other cases. Finally, to extensively support our conclusions, we conducted an additional test on a larger, but non subjectively annotated, dataset of video sequences that includes 19,840 PVSs. Here we defined an incoherence index as the the standard deviation of the different VQMs for each PVS, and concluded that the proposed *improved* VQMs seem to be superior because they return a lower incoherence index value.

The paper is organized as follows. In Section 2 related work is presented. Then, Section 3 presents the datasets used in this work, followed by an analysis of the behavior of some VQMs in Section 4. Our proposal is presented in Section 5 followed by experimental results in Section 6. Conclusions are drawn in Section 7.

2. RELATED WORK

Traditionally, two main approaches have been adopted for video quality assessment. The first one consider the possibility of predicting the perceived quality by modelling the properties of the human vision system (HVS) [2, 3, 4, 5]. The other approach instead focuses on determining how pronounced is the presence of artifacts by extracting features from the PVS and eventually also from the source and derive on the basis of these information the perceived quality [6]. Mapping features to perceived visual quality has been done in different ways, for instance using a weighted combination of the features [7, 8], through a ratio between information extracted from the PVS and the SRC to measure how distorted the PVS is [9] or exploiting machine learning algorithms.

The machine learning approach has demonstrated to be quite effective and hence has gained popularity in the last decade with the publication of a large number of works. In [10, 11, 12, 13] the authors extracted features directly at the pixel level and relied on machine learning algorithms to predict the quality of the PVS thus designing from scratch new VQMs. Recently other authors [14, 15] instead focused on fusing together some of existing VQMs using machine learning algorithms to derive better predictors of subjective quality.

The aim of this work is not to design a highly sophisticated NN for quality assessment as it has been largely done in literature but rather to propose a simple and thus computationally efficient NN based tool that is able to address a shortcoming of widely used and trusted VQMs. Furthermore, to the best of our knowledge, few works focused on trying to characterize the shortcomings of VMAF [16] that is generally regarded as a quite accurate VQM. In this work we objectively characterize PVSs for which even the VMAF is likely to inaccurately predict their MOS.

This work has been supported in part by the Politecnico Interdepartmental Center for Service Robotics (PIC4SeR, http://pic4ser.polito.it). Some of the computational resources were provided by HPC@POLITO (http://www.hpc.polito.it).

3. DATASETS DESCRIPTION

In this work we rely on the recently released *ITS4S* dataset, which has been assembled and published by NTIA/ITS [17] in the CDVL repository [18] to design and cross validate the proposed VQM improvement. It includes 813 unique source sequences at 1280x720 resolution, each about 4 second long. A subset comprising 514 of these sequences has been AVC compressed by the authors at one of these 5 different bitrate values: 512, 951, 1256, 1732, 2340 kbps. For lower bitrate encodings, lower resolutions have been used [19], however all content has been decoded and upscaled again at 1280x720 before performing any subjective evaluation. We run full-reference objective measures on all the available 514 SRC-PVS pairs[1], computing the following 5 objective measures: PSNR, SSIM, MS-SSIM, VIF as implemented by the VQMT software [20], and the VMAF v. 0.6.2 measure [1].

Our study also considers the Netflix Public Dataset [21], which includes 70 subjectively annotated PVSs covering the full MOS range and the VQEG JEG-Hybrid Large Scale dataset (JEG-DB) [22] which includes 19,840 PVSs obtained by compressing a few source sequences in HEVC format using a large variety of coding parameters, including bitrates ranging from 500 Kbps to 16 Mbps. Both datasets include high resolution content (1920x1080).

4. PRELIMINARY VQMS ACCURACY INVESTIGATION

In order to be able to analyse the prediction accuracy of the different VQMs on the annotated datasets, i.e., the ITS4S and the Netflix public datasets, we first fitted the values of all the 5 VQMs to the MOS using the following logistic function

$$\widehat{MOS} = \beta_1 \left(0.5 + \frac{1}{1+\exp\beta_2(VQM-\beta_3)}\right) + \beta_4 \cdot VQM + \beta_5 \quad (1)$$

as recommended in [23]. The five coefficients for each VQM are computed by performing a least square fitting on the ITS4S dataset.

Then we computed the Pearson correlation coefficient (PCC), the Spearman rank correlation coefficient (SRCC), and the residual mean square error (RMSE) to quantitatively compare the prediction accuracy of the different VQMs.

In this step we introduce two features related to the complexity of the scene and the sensitivity of the HVS with respect to spatial and temporal constraint. They are computed on each PVS, derived from the definition of the spatial and temporal perceptual information in ITU [24]:

$$AvgSI = \left(\sum_{i=1}^{N_F} std(Sobel(F_i))\right)/N_F \quad (2)$$

$$AvgTI = \left(\sum_{i=2}^{N_F} std(F_i - F_{i-1})\right)/N_F \quad (3)$$

where F_i, N_F and $std()$ denote respectively the i-th frame, the number of frames and the standard deviation.

To characterize the effect of the AvgSI and the AvgTI on the performance of the VQMs, we analyzed the three performance measures (PCC, SRCC, RMSE) separately for PVSs derived from sources characterized by different values of AvgSI and AvgTI. While we were not able to identify a trivial relation between the accuracy of the VQMs and the AvgTI, we noticed that the prediction accuracy of all the VQMs is clearly affected by low values of AvgSI.

[1]The dataset of the newly computed measures is publicly available for research purposes at http://media.polito.it/its4s

Table 1. VQMs prediction of MOS on the ITS4S dataset.

Index	PVS	PSNR	SSIM	MS-SSIM	VIF	VMAF
PCC	AvgSI<40	0.310	0.370	0.451	0.501	0.599
	AvgSI≥40	0.655	0.740	0.770	0.822	0.833
	All	0.593	0.688	0.731	0.772	0.799
SRCC	AvgSI<40	0.324	0.231	0.379	0.495	0.596
	AvgSI≥40	0.711	0.716	0.786	0.809	0.826
	All	0.619	0.624	0.707	0.747	0.777
RMSE	AvgSI<40	0.571	0.597	0.555	0.562	0.505
	AvgSI≥40	0.591	0.507	0.478	0.430	0.416
	All	0.585	0.528	0.496	0.462	0.437

Table 2. VQMs prediction of MOS on the Netflix public dataset.

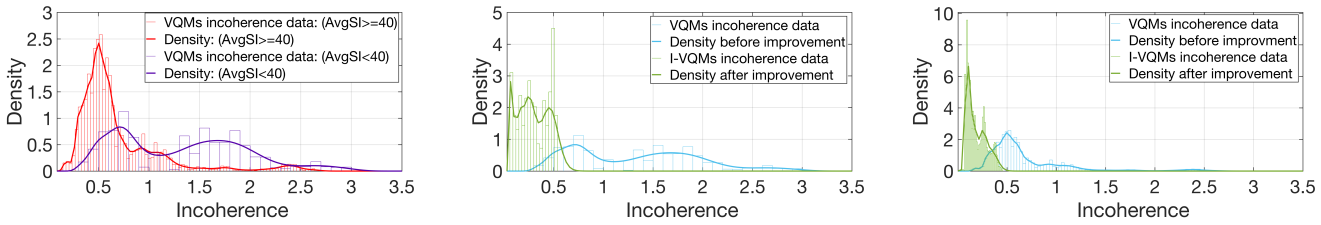
Index	PVS	PSNR	SSIM	MS-SSIM	VIF	VMAF
PCC	AvgSI<40	0.533	0.425	0.477	0.497	0.844
	AvgSI≥40	0.809	0.801	0.840	0.920	0.967
	All	0.656	0.669	0.735	0.783	0.935
SRCC	AvgSI<40	0.579	0.483	0.549	0.591	0.840
	AvgSI≥40	0.870	0.805	0.930	0.925	0.949
	All	0.682	0.654	0.769	0.779	0.922
RMSE	AvgSI<40	1.023	0.981	1.008	0.941	0.737
	AvgSI≥40	0.952	0.983	0.849	0.813	0.624
	All	1.001	0.982	0.904	0.857	0.663

In Table 1 and 2 we report, for the ITS4S and the Netflix dataset respectively, the PCC, SRCC and RMSE measures computed separately for PVSs derived from sources with low AvgSI (Group 1: AvgSI<40) and high AvgSI (Group 2: AvgSI≥40). The value 40 has been experimentally determined as the highest value at which there is still significant difference between the two groups. It can be clearly seen that the prediction of the VQMs poorly correlates with the MOS when evaluating the quality of PVSs in Group 1. Furthermore, higher RMSE values are observed, in general, for PVSs in Group 1.

In order to further investigate the performance of the VQMs on sequences having low AvgSI we carried out an additional analysis on the JEG-DB dataset. Despite the fact such dataset is not subjectively annotated, it contains a large number of PVSs, which helps in making observations more statistically significant. On this dataset we computed for each PVS the five objective VQMs and to each of them we applied the logistic functions in Eq. (1) to obtain the five MOS estimations as we did for the ITS4S and Netflix datasets.

We define a measure, which we name incoherence of the VQMs, by evaluating, for each PVS, the standard deviation of its five VQMs. In other words, we associate a numerical value I_{PVS} to each PVS that quantifies how incoherent are the predictions of the considered VQMs when used to estimate a MOS value. Aiming at studying the impact of low AvgSI on the value of I_{PVS}, we performed a non parametric estimation of the probability density function of the values of I_{PVS} both for PVSs of Group 1 and 2. Such probability density functions have been fitted from the I_{PVS} values using the kernel density estimation method, choosing Normal kernels. Further information about the procedure are available in [25].

Figure 1a reports the densities obtained for each group of PVSs. The shape of the densities clearly suggests that the VQMs are more likely to provide incoherent estimations when used to predict the MOS of a PVS belonging to Group 1. In fact, while the density of the PVSs in Group 2 concentrate most of the probability in the range [0,1], the density associated to Group 1 is spread over much higher



(a) Group 1 and Group 2 (before the improvement) (b) Group 1 (AvgSI<40) (c) Group 2 (AvgSI≥40)

Fig. 1. Distribution of the incoherence of the VQMs on the JEG-DB before (VQMs) and after the improvement (I-VQMs).

Table 3. Accuracy of MOS predicted by the proposed I-VQMs on the ITS4S dataset.

Index	PVS	I-PSNR	I-SSIM	I-MS-SSIM	I-VIF	I-VMAF
PCC	AvgSI<40	0.681	0.648	0.668	0.736	0.675
	AvgSI≥40	0.802	0.764	0.820	0.838	0.853
	All	0.788	0.751	0.802	0.824	0.829
SRCC	AvgSI<40	0.646	0.619	0.667	0.700	0.685
	AvgSI≥40	0.790	0.744	0.802	0.823	0.841
	All	0.771	0.724	0.780	0.802	0.812
RMSE	AvgSI<40	0.451	0.469	0.453	0.421	0.464
	AvgSI≥40	0.455	0.483	0.430	0.409	0.395
	All	0.454	0.480	0.435	0.411	0.411

Table 4. Accuracy of MOS predicted by the proposed I-VQMs on the Netflix dataset.

Index	PVS	I-PSNR	I-SSIM	I-MS-SSIM	I-VIF	I-VMAF
PCC	AvgSI<40	0.743	0.606	0.632	0.614	0.846
	AvgSI≥40	0.931	0.795	0.935	0.894	0.969
	All	0.876	0.743	0.857	0.822	0.937
SRCC	AvgSI<40	0.811	0.671	0.702	0.637	0.878
	AvgSI≥40	0.943	0.821	0.948	0.914	0.941
	All	0.899	0.761	0.871	0.815	0.926
RMSE	AvgSI<40	0.859	0.981	0.934	0.915	0.799
	AvgSI≥40	0.713	0.939	0.751	0.807	0.650
	All	0.765	0.960	0.816	0.844	0.702

incoherence values and reveals that the VQMs provide, for many sequences, MOS estimations that differ, among them, more than 1 unit of MOS.

In light of these results, it seems clear that low values of AvgSI strongly affects the performance of the VQMs. In the next section we will propose two strategies to improve the accuracy of the VQMs on the basis of this observation.

5. PROPOSED VQMS IMPROVEMENT

We propose to improve the VQMs using models that predict the value of the MOS considering, in addition to the value of the VQM itself, the AvgSI in order to overcome the shortcoming pointed out by the previous analysis and the AvgTI that allows to take into consideration the motion masking effects, i.e., lower perception of distortion in presence of large motion. Similar spatiotemporal features have been widely used while evaluating subjective quality [16, 14].

We designed and trained five NNs, one for each VQM. Each NN computes $\widehat{MOS}$, i.e., an improved estimation of the value of the MOS given the VQM, AvgSI, and AvgTI values of the PVSs. More formally, we have

$$I\text{-}VQM = \widehat{MOS} = NN_{VQM}(VQM, AvgSI, AvgTI) \quad (4)$$

where *I-VQM* and NN_{VQM} are respectively the improved VQM and the NN that improves the considered VQM. Exploiting the initial accuracy of each VQM coupled with the outcomes of the analysis presented in the previous section, we designed really shallow NNs with few features in input thus avoiding overfitting while minimizing computational costs. The best results have been obtained adopting a NN structure with three neurons on the input layer, one for each input, a single hidden layer with five neurons and a single neuron output layer that delivers the MOS prediction. All the 5 NNs have been trained on the ITS4S dataset using the Levenberg-Marquardt algorithm with backpropagation.

As it will be seen in the next section, each I-VQM is able to predict the quality of PVSs generated from sources characterized by low AvgSI more accurately than the corresponding VQM. However a not negligible gap in accuracy can still be observed when comparing the performance of the I-VQMs on the low and high AvgSI groups. To overcome this limitation of the I-VQMs we designed an additional NN that takes as input all the five VQMs, the AvgSI and the AvgTI, and produces, as output, an estimation of the MOS named VQMFI, i.e., VQM fused index. Except for the number of neurons on the input layer that we changed from three to seven, we maintained the structure of the NN similar to the one previously described.

At the end of our analysis, we have six trained NNs, i.e., five NNs that implement the improved version I-VQM of each VQM used in our study and another one that implements our proposed VQMFI.

6. EXPERIMENTAL RESULTS

In the experiments, we first perform a leave-one-out cross validation of the proposed NNs on the ITS4S dataset, then we perform additional tests on the Netflix dataset and the JEG-DB dataset that have never been involved in the training process.

6.1. I-VQMs and VQMFI performance on the ITS4S dataset

Comparing the values in Table 1 to those in Table 3, it appears evident that our proposal contributes to improve all the VQMs considered in this study since a significant gain in terms of correlation can be observed especially for the PVSs in Group 1. For instance, an increase of the PLCC from 0.31 to 0.68 for the PSNR is obtained. The results are further refined when using the VQMFI as it can be seen in Table 5.

Table 5. Performance of the proposed VQMFI compared, as a reference, to VMAF and our I-VMAF.

		ITS4S dataset			Netflix dataset		
PVS		VMAF	I-VMAF	**VQMFI**	VMAF	I-VMAF	**VQMFI**
PCC	AvgSI<40	0.599	0.675	**0.828**	0.844	0.846	**0.897**
	AvgSI≥40	0.833	0.853	**0.854**	0.967	0.969	**0.972**
	All	0.799	0.829	**0.844**	0.935	0.937	**0.951**
SRCC	AvgSI<40	0.596	0.685	**0.822**	0.840	0.878	**0.902**
	AvgSI≥40	0.826	0.841	**0.844**	0.949	0.941	**0.968**
	All	0.777	0.812	**0.840**	0.922	0.926	**0.950**

Table 6. Testing the superiority of VQMFI in terms of PLCC with respect to VMAF on the PVSs in Group 1.

Dataset	ITS4S	NETFLIX
p-value	0.001	0.153

Table 7. Quantitative statistics of the VQMs and I-VQMs incoherence.

	VQMs		I-VQMs	
Index	SI<40	SI≥40	SI<40	SI≥40
Minimun	0.32	0.13	**0.06**	**0.04**
First quartile	0.77	0.44	**0.17**	**0.10**
Average	1.31	0.69	**0.31**	**0.19**
Third quartile	1.79	0.75	**0.46**	**0.26**
Maximum	2.98	2.70	**0.52**	**0.45**

6.2. I-VQMs and VQMFI performance on the Netflix dataset

We then used the six NNs, trained using all the 514 PVSs in the ITS4S dataset, to predict the MOS of the PVSs in the Netflix dataset and then compute the correlations coefficients as previously done. Results are shown in Table 4 for all the I-VQMs and in Table 5 for the VQMFI. Also in this case our proposal shows better accuracy than all the other VQMs, including VMAF which was developed by Netflix and thus probably trained also on the Netflix dataset itself. It is worth nothing that our proposal performs significantly better than other VQMs when considering PVSs having low value of SI, which is the most difficult condition as observed in the first part of this work.

6.3. I-VQMs performance on the JEG-DB dataset

The experiments conducted on the JEG-DB dataset aimed at demonstrating that the proposed I-VQMs lead to less incoherent MOS estimations than the original VQMs. To this end, we used our five *NN*s to compute the five I-VQMs for each PVS in the JEG-DB dataset. Then, in order to compare the incoherence of the VQMs before and after the improvement, we computed again the I_{PVS} for each PVS as done before, but this time considering the MOS estimations provided by the I-VQMs. Hence we fitted the probability distribution of the incoherence values after the improvement as done previously. Figure 1 shows the densities of the incoherence of the VQMs with and without the proposed improvement for the PVSs in Group 1 and 2. It can be easily noticed that the I-VQMs provide MOS estimations which are much more coherent than the original VQMs as the corresponding densities have supports [0, 0.5] while the densities associated to original VQMs suggest that that MOS estimations are expected to differ for more than 0.5 with high probability. The values in Table 7 are reported to provide a quantitative representation of the performance of the proposed improvements. It can be seen that the I-VQMs are expected to yield less incoherent estimations of the MOS since all the statistical indicators of incoherence after the improvement are clearly outperformed by those observed before the proposed improvement.

6.4. Statistical tests

We conducted two Z-tests to assess the statistical significance of the superiority of the VQMFI with respect to the VMAF and consequently to the other VQMs in terms of Pearson correlation to the MOS on the ITS4S and Netflix public dataset. From the results reported in Table 6, it seems that for the PVSs in Group 1, the VQMFI is expected to provide subjective quality estimation that correlates to the MOS better than that of all the metrics considered in this work with more than 84% (1−p-value) of confidence.

7. CONCLUSIONS

In this work we focus on investigating the performance of well-known objective video quality metrics in order to improve their accuracy. Relying on three datasets, we observed that such accuracy is lower when sources are characterized by low average spatial activity. Then, we proposed a machine learning based improvement of each of the considered VQMs, as well as a video quality metrics fusion index (VQMFI) obtained by jointly exploiting the VQMs, the average spatial and temporal activity indexes. The results demonstrate the ability of the VQMFI to accurately predict the subjective score of PVSs both for low and high spatial activity values. Different datasets have been used to validate the results.

8. REFERENCES

[1] Netflix, "VMAF - video multi-method assessment fusion," https://github.com/Netflix/vmaf, Jan. 2019.

[2] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, no. 2, pp. 231–252, 1999.

[3] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," in *Digital Video Compression: Algorithms and Technologies 1996*. International Society for Optics and Photonics, 1996, vol. 2668, pp. 450–462.

[4] S. Winkler, "Perceptual distortion metric for digital color video," in *Human Vision and Electronic Imaging IV*. International Society for Optics and Photonics, 1999, vol. 3644, pp. 175–185.

[5] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet, "A metric for no-reference video quality assessment for hd tv delivery based on saliency maps," in *2011 IEEE International Conference on Multimedia and Expo*, July 2011, pp. 1–5.

[6] S. Süsstrunk and S. Winkler, "Color image quality on the internet," *Proc. SPIE Electronic Imaging 2004: Internet Imaging V*, vol. 5304, pp. 118–131, 2004.

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

[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The 37th Asilomar Conference on Signals, Systems Computers*, Nov. 2003, vol. 2, pp. 1398–1402.

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

[10] P. Le Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1316–1327, Sep. 2006.

[11] P. Gastaldo, S. Rovetta, and R. Zunino, "Objective assessment of MPEG-video quality: a neural-network approach," in *Proc. International Joint Conference on Neural Networks*, July 2001, vol. 2, pp. 1432–1437 vol.2.

[12] M. Narwaria and W. Lin, "SVD-based quality metric for image and video using machine learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 347–364, April 2012.

[13] C. Wang, X. Jiang, F. Meng, and Y. Wang, "Quality assessment for MPEG-2 video streams using a neural network model," in *13th International Conference on Communication Technology*. IEEE, 2011, pp. 868–872.

[14] J. Y. Lin, T. Liu, E. C. Wu, and C. . J. Kuo, "A fusion-based video quality assessment (FVQA) index," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec 2014, pp. 1–5.

[15] V. Lukin, N. Ponomarenko, O. Ieremeiev, K. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9394, 03 2015.

[16] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[17] "National telecommunications and information administration (NTIA), institute for telecommunication sciences (ITS)," 2019.

[18] CDVL Technical Committee, "The consumer digital video library," http://www.cdvl.org, 2019.

[19] M. H. Pinson, "ITS4S: A video quality dataset with four-second unrepeated scenes," *NTIA Technical Memo TM-18-532*, Feb. 2018.

[20] P. Hanhart and R. Hahling, "Video quality measurement tool (VQMT)," http://mmspg.epfl.ch/vqmt, Sept. 2013.

[21] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Data Compression Conference (DCC)*, April 2017, pp. 52–61.

[22] M. Barkowsky, E. Masala, G. Van Wallendael, K. Brunnström, N. Staelens, and P. Le Callet, "Objective video quality assessment-towards large scale video database enhanced model development," *IEICE Transactions on Communications*, vol. E98B, no. 1, pp. 2–11, 2015.

[23] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, June 2011.

[24] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," Apr. 2008.

[25] B. A. Turlach, "Bandwidth selection in kernel density estimation: A review," in *CORE and Institut de Statistique*, 1993.