Analyzing the effect of adding temporal features to an autoencoder-based video quality model

André H. M. Costa[1], Helard Becerra Martinez[2], Daniel G. Silva[1], and Mylène C.Q. Farias[1]

[1]Department of Electrical Engineering, University of Brasília, Brazil.
[2]University College Dublin, Ireland.

Abstract

According to Cisco, most Internet traffic is currently comprised of videos. Therefore, developing a quality assessment method for assuring that those videos are received and displayed with quality at the user side is an important and challenging task. As a consequence, over the last decades, several no-reference video quality metrics have been proposed with the goal of blindly predicting (with no access to the original signal) the quality of videos in streaming applications. One of such metrics is NAVE, whose architecture includes an auto-encoder module that produces a compact set of visual features with a higher descriptive capacity. Nevertheless, the visual features in NAVE do not include descriptive temporal features that are sensitive to temporal degradation. In this work, we analyze the effect on accuracy performance of using a new type of temporal features, based on natural scene statistics. This approach has the goal of making the tested video quality metric more generic, i.e. sensitive to both spatial and temporal distortions and therefore adequate for video streaming applications.

Introduction

Quality metrics can be divided into three classes: full-reference (FR), where the quality assessment uses the reference data, reduced reference (RR), where only part of the reference information is used, and no-reference (NR), where no information about the reference is available. In video streaming applications, the reference content is not frequently available and, therefore, NR metrics are the most suitable choice. Although the area of video quality assessment (VQA) has evolved a lot in the last decades, there are few studies about the effect of temporal degradation on the overall video quality. In fact, most VQA metrics use image quality assessment (IQA) metrics to estimate the quality of each frame and, then, combine the frame predictions to obtain an overall quality prediction. As expected, these approaches are not able to satisfactorily capture common temporal degradations, such as frame freezing and packet-loss. Among the works that have used temporal features to estimate video quality, a recent work by Sinno and Bovik [1] proposes temporal features, which are based on natural scene statistics (NSS), has shown good results in previous studies [2, 3].

This study explores the usage of a refined set of temporal features, focusing on the identification of common temporal distortions, such as packet-loss and frame freezing. In a previous work, we designed a NR video quality metric, NAVE, that uses an autoencoder architecture [4]. A following study explored the performance of this model by using different architecture setups

[5]. One scenario that was not yet covered is the impact of temporal features on the overall performance of the metric. We propose to incorporate a set of descriptive temporal features, which are capable of detecting temporal degradation, and evaluate their impact on the performance of NAVE.

NAVE Metric

NAVE is a VQA metric composed of 2 autoencoder layers and a classification function [4]. In this work, we excluded the SI and TI [6], spatio-temporal features, while the spatial video features, based on NSS and extracted as in the image quality metric DIIVINE [7], were kept. The autoencoders module reduces the input vector into a low-dimension set of features. The first autoencoder layer receives the full input feature vector of size $m \times n$, where m is the number of features and n is the number of frames, and reduces it to a low-dimension vector of size $50 \times n$ - AE1. The second autoencoder layer takes the AE1 vector and further reduces its size, yielding a $20 \times n$ vector. The resulting vector AE2 is not only a reduced version of the original input, but is thought to contain partial information of all the input features. The AE2 vector serves as input for the classification module (softmax function), which returns a quality score for each single frame in the video clip. These quality scores are then averaged to obtain the overall video quality prediction. It is worth pointing out that the only temporal feature used in NAVE metric is the temporal index (TI) [6], which is the difference between consecutive frames in a video signal.

NAVE was trained using a subset of the UnB-AVQ database [8, 9] which is an audio-visual quality database with degradations on both video and audio components. In this work, we used UnB-AVQ-Experiment1 subset of the database, which is comprised of audiovisual sequences with video-only distortions: Bitrate compression, Packet-Loss and Frame-Freezing. The UnB-AVQ-Experiment1 dataset is comprised of 60 videos with spatial resolution of 1280x720, 30 fps of temporal resolution and 4:2:0 color distribution. All videos were compressed with H.264 and H.265 coding algorithms. Additionally, 5 degrees of Packet-Loss and Frame-Freezing distortions were generated separately. These distortions were combined with the different bitrate levels of compression and they resulted in 12 different test conditions.

Visual and Temporal Features

We introduced new features to the NAVE metric in order to improve its performance over temporal degradations such as Packet-Loss and Frame-Freezing. The first type of temporal feature is based on Sinno and Bovik's work [1]. These temporal

Video Under Test

Signal Feature Extraction

Deep Net Model

Autoencoder 1

Autoencoder 2

Classification

Softmax

Probabilities Matrix

Output Processing
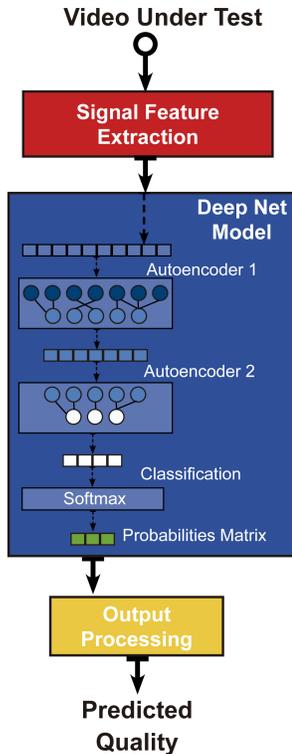
Predicted Quality

Figure 1: Architecture of the NAVE quality metric.

features are based on the fact that different distortions change the shape of the Gaussian distribution of the mean subtracted contrast normalized coefficients (MSCN) [2, 3]. These MSCN coefficients are computed by performing a shifted subtraction of consecutive frames. More specifically, the pixels $I_t(i,j)$ of the t-the frame are subtracted from the pixels $I_{t+1}(i,j-1)$, $I_{t+1}(i-1,j)$, $I_{t+1}(i-1,j-1)$ and $I_{t+1}(i-1,j+1)$ of the following frame, yielding four directional difference frames. Then, a 7×7 normalized Gaussian filter is used to compute the local average value of each difference frame $D_t(i,j)$. Finally, we obtain the MSCN coefficients by computing the Gaussian average of the directional difference frames and dividing the result by their variance. Afterwards, the frame of coefficients is divided into 96×96 patches, similarly to what is done in the work by Mittal *et al.*[2]. The patches of each frame are then fitted into a Generalized Gaussian Distribution (GGD), using the moment matching function proposed by Lasmar *et al.* [10]. Earlier results [2, 3] show that coefficients of pristine and distorted videos tend to have different distributions, hence, GGD coefficients can be used to predict picture quality.

The second type of features added in this work are the coefficients of the BRISQUE quality metric [3]. Similar to the approach proposed by Sinno and Bovik [1], BRISQUE coefficients are calculated using a NSS approach. The algorithm first computes the MSCN coefficients of each video frame. The MSCN is obtained by subtracting the pixel value (i,j) of the local average at (i,j) and dividing the result by the local variance at pixel (i,j). Then, these MSCN coefficients are fitted into a GGD [11]. Differently from the MSCN coefficients computed for the work of Sinno and Bovik [1], the BRISQUE MSCN coefficients are computed from the frame pixel values instead of the differences of subsequent

Table 1: Features sets combinations used as input to the autoencoder of NAVE.

Features set 1	DIIVINE
Features set 2	BRISQUE + SinnoP
Features set 3	DIIVINE + BRISQUE
Features set 4	DIIVINE + SinnoM
Features set 5	DIIVINE + SinnoP
Features set 6	DIIVINE + BRISQUE + SinnoM
Features set 7	DIIVINE + BRISQUE + SinnoP
Features set 8	SinnoP

frames. The local average is computed using a 7×7 normalized Gaussian filter and the variance is computed using a 7×7 filter window around the pixel. Additional coefficients are obtained by the product of neighboring coefficients located at $(i,j+1)$, $(i+1,j)$, $(i+1,j+1)$ and $(i+1,j-1)$ positions. These differences of coefficients follow an asymmetric generalized Gaussian distribution and are fitted using the moment matching function proposed by Lasmar *et al.* [10].

In summary, in this study, we used 3 sets of features: spatial features from DIIVINE, spatial features from BRISQUE, and temporal features from Sinno and Bovik [1]. Since the Sinno features produce GGD coefficients for a number of patches in each frame, we decided to use two pooling strategies to have a better understanding of their quality assessment capability. For the first strategy, we took the mean value for each coefficient in each frame, which resulted in a $8 \times n$ feature vector. This strategy is referred in this work as SinnoM. For the second strategy, we used each coefficient as a feature. Degradations affect different parts of the frame and, for that reason, coefficients from each patch may carry relevant information about the video quality, which would otherwise be lost in the averaging process. As a result, we generated a feature vector of size $184 \times n$ that is referred as SinnoP. No processing was made to the features of BRISQUE, thus, its original size of $36 \times n$ was maintained. No changes were made to the architecture of NAVE with respect to the output of the two autoencoder layers. This means that, although the original dimension ($90 \times n$) of the input vector changed, the dimension of the output vector ($20 \times n$) remained the same. The different combinations of the feature sets are presented in Table 1.

Performance Analysis

NAVE model was re-trained using as input all of the 8 sets of features presented in Table 1. We used a 10-fold cross validation procedure to train all 8 feature sets on NAVE's architecture. To do so, we extracted and combined all corresponding features from the UnB-AVQ-Experiment1 dataset[8]. Once extracted, a temporal sub-sampling was applied to ease the training process. The number of frames were averaged to a tenth of the total number of frames. Further details of the training procedure can be found in Martinez *et al.*' previous work [4].

The performance analysis was organized in three parts. First, we present and analyse the results obtained using the different feature sets displayed in Table 1. The objective is to understand the behaviour of the NAVE autoencoder architecture when new features are added. More specifically, we want to check which of the new features are able to improve the NAVE performance. In the second part of the analysis, we compare the model performance obtained with the best set of features with other image quality

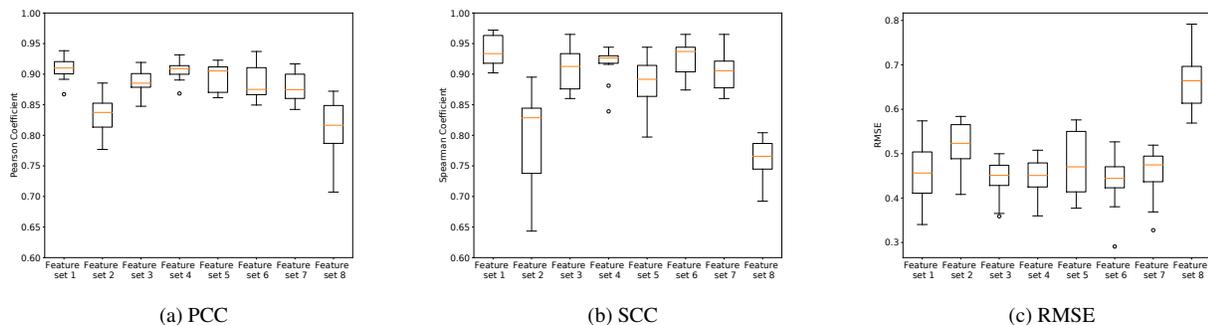| (a) PCC | (b) SCC | (c) RMSE |

Figure 2: Box Plots of PCC, SCC and RMSE for overall results of all 8 feature Sets (see Table 1).

Table 2: PCC, SCC, RMSE for the different feature combinations tested on the UnB-AVQ Database.

Features set	Measure	Packet-Loss	Freezing	All
Feature set 1	PCC	0.944	**0.910**	**0.909**
	SCC	**0.965**	0.920	**0.937**
	RMSE	0.426	0.477	0.457
Feature set 2	PCC	0.944	0.763	0.833
	SCC	0.834	0.822	0.796
	RMSE	0.424	0.588	0.515
Feature set 3	PCC	0.948	0.849	0.886
	SCC	0.902	0.914	0.907
	RMSE	0.386	0.486	0.442
Feature set 4	PCC	0.941	**0.909**	**0.905**
	SCC	0.937	**0.925**	0.915
	RMSE	0.424	**0.456**	**0.444**
Feature set 5	PCC	0.934	0.889	0.894
	SCC	0.885	0.891	0.886
	RMSE	0.456	0.493	0.477
Feature set 6	PCC	**0.951**	0.851	0.887
	SCC	0.942	0.920	0.925
	RMSE	**0.370**	0.489	0.436
Feature set 7	PCC	0.949	0.834	0.878
	SCC	0.914	0.897	0.904
	RMSE	0.388	0.507	0.454
Feature set 8	PCC	0.834	0.817	0.807
	SCC	0.805	0.765	0.761
	RMSE	0.591	0.724	0.662

metrics over the UnB-AVQ-Experiment1 Database. Finally, we compared the performance of the model for specific test conditions, taken from the UnB-AVQ-Experiment1 Database.

Feature Sets Performance

Table 2 presents the results obtained testing all feature sets combinations. This table depicts the average value Pearson and Spearman correlation coefficients (PCC and SCC) and the root mean squared error (RMSE) for the 10-fold cross-validation procedure. The average of these 3 performance metrics are reported for each type of degradation and for the overall test cases (all types of degradations in the database). It is worth pointing out that in the UnB-AVQ-Experiment1 Database the Packet-loss and Frame-Freezing appear in combination with compression distortions.

In Table 2, we notice that there are no major changes in the performance metrics results across the different features sets, specially for Packet-loss degradations. Results obtained with Feature sets 1, 4, and 6 were slightly better than what was obtained with the other feature sets. An unexpected result was the per-

formance of DIIVINE and Sinno features. First, it was surprising that Feature set 1, which only contains the DIIVINE features (spatial features), obtained good results for both Packet-loss and Frame-Freezing degradations. Second, it was surprising that the Sinno features had a smaller than expected contribution to the quality prediction performance for Packet-Loss degradations. More specifically, comparing Feature sets 1 and 5, it is possible to see a negative effect of the SinnoP features on the average correlation values. Furthermore, Feature set 8, which only contains the SinnoP features, was the worst performing feature set for the Packet-Loss degradations. On the other hand, when compared to the Feature set 1, the SinnoM features in Feature sets 4 and 6 did not negatively impact the correlation coefficients. Also, when concatenated with BRISQUE features in Feature set 6, these features improved both PCC and RMSE values. Although the performance improvement for Feature Set 6 cannot be credited to BRISQUE or SinnoM alone, these results show that there is an interaction between these two features.

For Frame-Freezing degradations, the features sets can be clustered into two different groups. The first group is comprised of Features sets 1 and 4, which show a better performance than the remaining features sets. Among the feature sets with the worst results is the Feature set 2, which contains the features from the DIIVINE and BRISQUE metrics. It can be inferred that the decrease in quality assessment is due to the BRISQUE insensitivity to Frame-freezing degradations. This effect can also be observed when comparing results of Feature sets 4 and 6. Both these sets contain SinnoM features, but only set 6, which is the worst performing set, has BRISQUE features. Feature Set 1 has a slightly better performance, which is expected since the DIIVINE feature vector is 11 times larger than the SinnoM feature vector. Therefore, the good performance of the DIIVINE features alone is intriguing. Similarly to what happened for Packet-Loss, the SinnoP features have an inferior performance when compared to the SinnoM features. This seems to indicate that the mean coefficients of each frame already carry most of the relevant information.

For better visualization, the distribution for the overall results are presented in Figure 2. This figure depicts box-plots of the PCC, SCC and RMSE values obtained for the test sequences of UnB-AVQ-Experiment1 Database. The distribution of the PCC values depicted in Figure 2 (a) show that the NAVE model performs best when features sets 1 and 4 are used as input. The median PCC values for these two feature sets are higher than for the others feature sets. Also, the lower limits of these distributions

Table 3: PCC, SCC, RMSE for the different feature combinations tested on the UnB-AVQ Database.

Features set	Measure	Packet-Loss	Freezing	All
PSNR	PCC	0.363	0.661	0.310
	SCC	0.434	0.600	0.314
	RMSE	13.623	8.664	11.436
SSIM	PCC	0.103	0.578	0.157
	SCC	0.125	0.577	0.163
	RMSE	2.364	2.852	2.621
DIIVINE	PCC	-0.889	-0.89	-0.858
	SCC	-0.862	-0.85	-0.855
	RMSE	2.515	2.881	2.705
VIIDEO	PCC	-0.578	-0.517	-0.542
	SCC	0.571	-0.440	-0.500
	RMSE	2.304	2.685	2.502
BIQI	PCC	-0.756	-0.576	-0.692
	SCC	-0.731	-0.640	-0.694
	RMSE	25.380	24.532	24.326
NIQE	PCC	-0.569	0.808	-0.702
	SCC	-0.634	0.828	-0.714
	RMSE	2.226	2.109	2.179
BRISQUE	PCC	-0.772	-0.887	-0.835
	SCC	-0.817	-0.908	-0.866
	RMSE	2.226	41.726	43.367
NAVE	PCC	0.933	0.895	0.896
	SCC	**0.942**	0.914	**0.917**
	RMSE	0.428	0.499	0.468
NAVEv2	PCC	**0.941**	**0.909**	**0.905**
	SCC	0.937	**0.925**	0.915
	RMSE	**0.424**	**0.456**	**0.444**

were higher than the lower limits of the distributions of the other features sets, which means that in a worst case scenario these features obtained better results. Figures 2 (b) and (c) present the distribution for the SCC and RMSE values. The results presented in these figures show that the feature set 4 takes the lead as the best performing set. Despite the fact that other features sets have a higher upper-limit for the distribution, feature set 4 has a tighter distribution and high median value. Overall, feature set 4 has slightly better results when compared to the Feature set 1. Based on these results, we used Feature set 4 as the autoencoder input features for the NAVE model. In the following sections, we refer to this variation of the NAVE metric as NAVEv2.

Comparison with Other Quality Metrics

We compare the NAVEv2 performance with the performance of the following popular FR and NR quality metrics:

- FR IQA metrics: SSIM [12] and PSNR;
- NR IQA metrics: DIIVINE [7], BIQI [13], NIQE [2], and BRISQUE [3];
- FR VQA metrics: VIIDEO [14] and NAVE [4].

All the metrics were tested using a 10-fold cross-validation training and testing methodology. Table 3 depicts the PCC, SCC and RMSE average values for different FR and NR quality metrics. Results are organized by degradation and overall score, with the best values highlighted in bold. Notice that the NAVEv2 metric is the best performing video quality assessment metric for the UnB-AVQ-Experiment1 Database, i.e. it achieves the highest PCC and SCC values and the lowest RMSE values. The distributions of the PCC and SCC values, presented in the box-plots in Figure 3, have much smaller variances than what is achieved by other metrics.
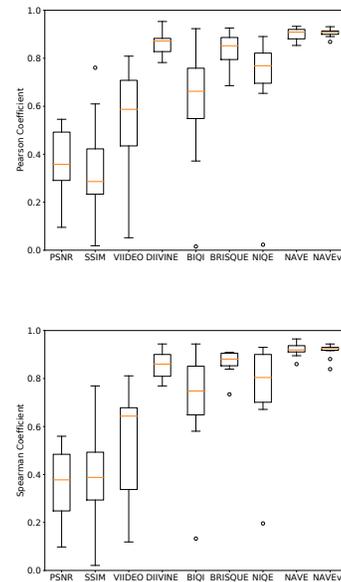

Figure 3: Pearson and Spearman coefficients for the FR and NR metrics trained and tested on the UnB-AVQ database.

The tighter distributions shows that even when the lowest performance values of NAVEv2 still perform better than other metrics.

Temporal and Compression Degradations

Finally, we analyse the performance of NAVEv2 across different contents, compression codecs, and temporal degradations. With this goal, we studied NAVEv2 performance for a couple of test conditions of the UnB-AVQ-Experiment 1 database. Figures 4 and 5 depict scatter plots of the quality predictions versus the mean opinion score (MOS), grouped by Packet-Loss and Frame-Freezing and the compression codec algorithm. More specifically, Figure 4 (a) and (b) show the scatter plots for all test conditions using H.264 codec. Notice that there is a clear difference in the predictions for Packet-Loss and Frame-Freezing scenarios. From Figure 4 (a), we observe that predictions for test conditions HRC1 (H.264, Bitrate = 500kb/s, PLR=10%) and HRC3 (H.264, Bitrate = 2000kb/s, PLR=5%) overlapped, meanwhile, ANC1 (no degradtions) predictions grouped separately. For Frame-Freezing in Figure 4 (b), the HRC7 (H.264, Bitrate = 800kb/s, N=3 freezing events at 3 different positions with lengths L = 2s and 3s), HRC9 (H.264, Bitrate = 2000kb/s, N=2 freezing events at 2 different positions with lengths L = 1s and 3s) and HRC10 (H.264, Bitrate = 200kb/s, N=1 freezing events at the begining of the video with length of L = 2s). Notice that these conditions were better clustered and aligned with the diagonal of the plot. It is worth pointing out that the UnB-AVQ database has a high content diversity, which might explain the high vertical spread in the MOS axis. In other words, the different contents are differently affected by the distortion and the quality prediction are spread out.
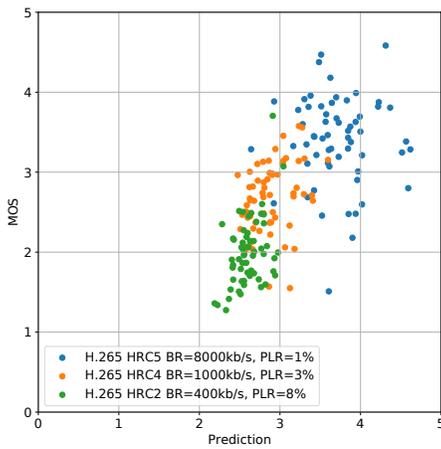
This behaviour also affects the videos encoded with the H.265 standard, as seen in Figure 4 (c) and (d). In this figure, we see an opposite behaviour with respect to each degradation scenario. Now, the best performance happens for Packet-Loss
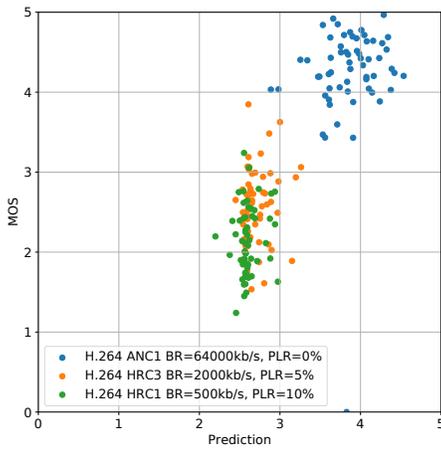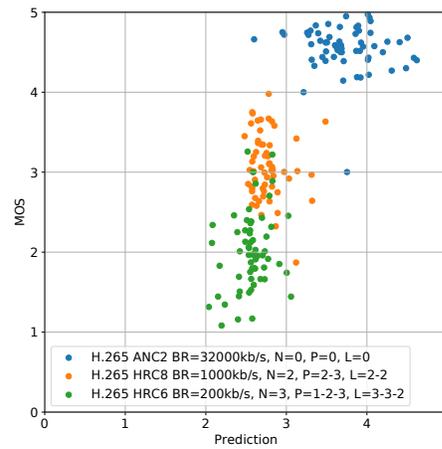
(a) H.264 + Packet-Loss
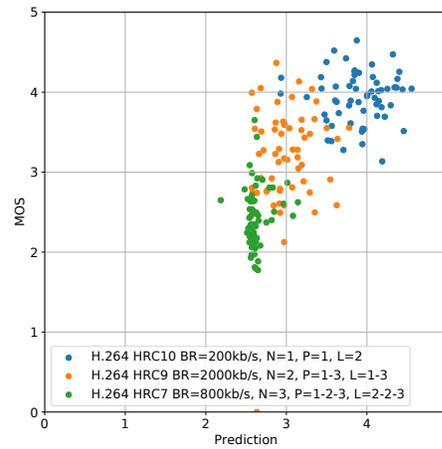
(b) H.264+ Frame-Freezing

(c)H.265 + Packet-Loss

(d) H.265 + Frame-Freezing

Figure 4: Scatter plot of the videos of the UnB-AVQ database compressed with H.264 and H.265 codecs, separated by Packet-Loss and Frame-Freezing degradations.

degradation with a better distribution around the diagonal of the plot and a better separation between degradation scenarios. The Frame-Freezing test conditions, HRC8 and HRC6, have prediction scores with similar values. This effect is worse for the group of videos with the largest amount of degradations, HRC6 and HRC7. In these test conditions, similar quality scores were attributed to most video sequences.

Figure 5 shows the scatter plot of the quality predictions versus the mean observer scores (MOS) for test conditions HRC3 and HRC4 (Packet-loss) and HRC8 and HRC9 (Frame-freezing). These plots contain tests conditions where the effect of bitrate compression is equivalent, which allows comparing the results regardless of the compression artifacts and, therefore, checking to see if our model is able to detect the temporal degradations. In Figure 5 (a), test condition HRC4 (orange) has a packet-loss ratio of 3% and HRC3 (blue) has a packet-loss ratio of 5% . This smaller levels of degradation is evident in the plot, since the distri-

bution of the orange points have both higher values of prediction score and MOS. There is a big vertical spread of the points for each case, which is attributed to how the different video contents are affected by the degradations. Figure 5 (b) show the results for test conditions HRC8 and HRC9, with Frame-Freezing degradations. Test condition HRC8 had the freezing pauses inserted in the middle and at the end of the video sequence, while the HRC9 had the pauses at the beginning and in the middle of the video sequence. It is known that events in the begining of the video are less annoying to viewers than events in the middle or beginning of the video [15]. This is evident in Figure 5 (b) where lower prediction scores and MOS values were reported for HRC8. The mean values of each scenario are found near the main diagonal of the plot, indicating a good quality assessment accuracy.
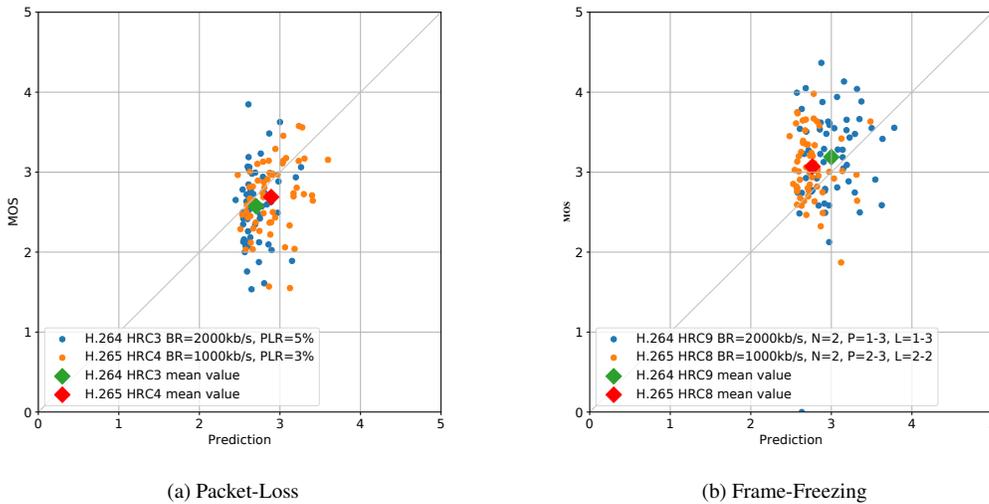
| | (a) Packet-Loss | (b) Frame-Freezing |

Figure 5: Scatter plot of the videos of the UnB-AVQ database compressed with H.265 separated by Packet-Loss and Frame-Freezing degradations.

Discussion

Video Streaming grew in popularity over the last decade and so grew the occurrence of temporal degradation. Temporal degradations have not been explored as much as spatial degradations due to its recent history. The results shown in the previous sections raised some points about the quality of videos containing temporal degradations. First, it is evident that the good performance of NSS-based features depends on their sensitivity to temporal degradations. Image quality metrics like DIIVINE and BRISQUE, which were originally designed to identify spatial degradations, presented a good performance for temporal degradations. Our initial hypothesis was that features from the DIIVINE metric, which were used in the original NAVE metric, captured partially the temporal degradations and, therefore, the addition of Sinno features would increase the metric performance. Nevertheless, our results showed that features from the DIIVINE metric were able to capture temporal degradations and only a small improvement was achieved with the addition of SinnoM features. Second, the metric was able to predict the different levels of Packet-Loss and Frame-Freezing degradations, even when the video was compressed. Since the H.265 codec requires a smaller bitrate than the H.264, the quality of the videos compressed with H.265 is going to be more impacted by packet-loss. On the other hand, the higher bitrate required by H.264 makes the frame-freezing events stand out. We believe that the interaction between the bitrate compression and the other temporal degradations needs to be further explored in other databases with more specific tests conditions.

Conclusions

The main focus of this work was to improve the NAVE metric at assessing the quality of videos with temporal degradations and better understand how these degradations affect the quality perception. We first tested different combinations of previously used features and new features. Among the features sets, the combination of DIIVINE and SinnoM features had the best perfor-

mance, improving the correlation between the prediction scores and the database MOS values. We tested the metric NAVEv2 with different scenarios of the UnB-AVQ database and found that for scenarios where the bitrate of the two codecs provided comparable visual quality, the metric was able to detect the different levels of degradations. Results found in this work bring up many future work possibilities, including the studying the contribution of combinations of compression, packet-loss, and frame-freezing degradations to visual quality.

References

[1] Z. Sinno and A. C. Bovik, "Spatio-temporal measures of naturalness," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1750–1754, IEEE, 2019.

[2] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[3] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[4] H. B. Martinez, M. C. Farias, and A. Hines, "A no-reference autoencoder video quality metric," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1755–1759, IEEE, 2019.

[5] H. Martinez, A. Hines, and M. C. Farias, "How deep is your encoder: an analysis of features descriptors for an autoencoder-based audio-visual quality metric," *arXiv preprint arXiv:2003.11100*, 2020.

[6] A. Ostaszewska and R. Kłoda, "Quantifying the amount of spatial and temporal information in video test sequences," in *Recent Advances in Mechatronics*, pp. 11–15, Springer, 2007.

[7] Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-diivine: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Signal Processing: Image Communication*, vol. 29, no. 7, pp. 725–747, 2014.

[8] H. B. Martinez, A. Hines, and M. C. Farias, "Unb-av: An audio-visual database for multimedia quality research," *IEEE Access*,

vol. 8, pp. 56641–56649, 2020.

[9] H. B. Martinez and M. C. Farias, "Full-reference audio-visual video quality metric," *Journal of Electronic Imaging*, vol. 23, no. 6, p. 061108, 2014.

[10] N.-E. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2281–2284, IEEE, 2009.

[11] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995.

[12] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402, Ieee, 2003.

[13] A. Moorthy and A. Bovik, "A modular framework for constructing blind universal quality indices," *IEEE Signal Processing Letters*, vol. 17, 2009.

[14] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2015.

[15] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Transactions 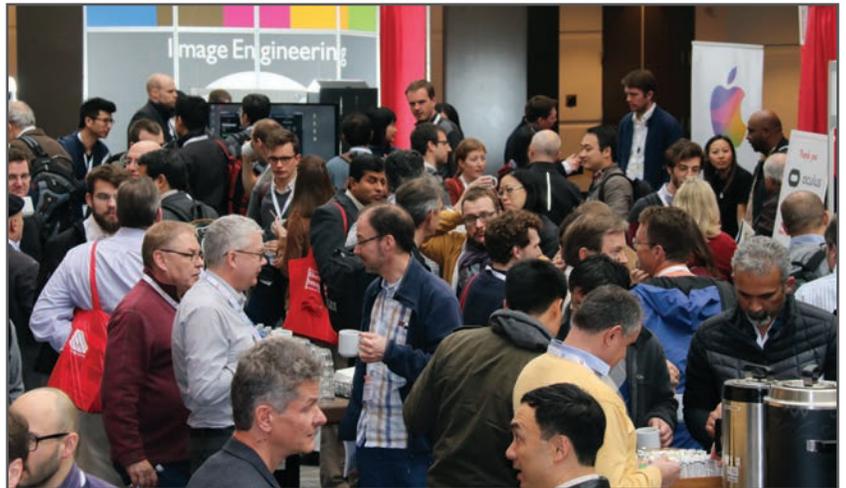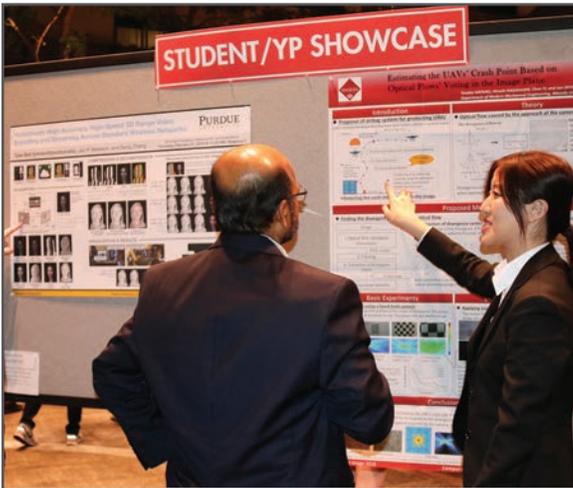on Image Processing*, vol. 26, no. 11, pp. 5217–5231, 2017.