Exploring the boundaries of an AE-based quality model: a performance analysis via synthetic content

Helard Becerra Martinez[2], André H. M. da Costa[1], Bruna Azambuja[1], Andrew Hines[2], and Mylène C.Q. Farias[1]

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

Abstract

The No-reference Autoencoder VidEo (NAVE) metric is a video quality assessment model based on an autoencoder machine learning technique. The model uses an autoencoder to produce a set of features with a lower dimension and a higher descriptive capacity. NAVE has been shown to produce accurate quality predictions when tested with two video databases. As it is a common issue when dealing with models that rely on a nested non-linear structure, it is not clear at what level the content and the actual distortions are affecting the model's predictions. In this paper, we analyze the NAVE model and test its capacity to distinguish quality monotonically for three isolated visual distortions: blocking artifacts, Gaussian blur, and white noise. With this goal, we create a dataset consisting of a set of short-length video sequences containing these distortions for ten very pronounced distortion levels. Then, we performed a subjective experiment to gather subjective quality scores for the degraded video sequences and tested the NAVE pre-trained model using these samples. Finally, we analyzed NAVE quality predictions for the set of distortions at different degradation levels with the goal of discovering the boundaries on which the model can perform.

Introduction

Similarly to other research areas, video quality assessment (VQA) has experienced a great amount of progress due to the adoption of several machine learning tools [1]. Several proposed ML-based quality models achieved very accurate quality predictions, however, in most cases, very little attention was paid to explaining the reasons behind these results. Being able to explain a model's output is not only important for the validation of the model itself, but it is also determinant for gaining new insights into the problem at hand. Moreover, any improvement of the model demands a good understanding of the model functionality [2, 3].

The NAVE metric is a No-reference video quality model based on an autoencoder technique [4]. The metric uses the autoencoder to produce a compact set of visual features, which have a higher descriptive capacity (see Figure 1). NAVE was tested using two video databases and it produced quality predictions with good correlations [5, 6]. Yet, it is not entirely clear at what level the content of the video sequences, or the actual visual distortions, are affecting the model's prediction. In order to further develop the model, exploration is required to better understand whether the model is robust when responding to visual impairments or if the quality predictions are biased and masked by the content itself.

This paper analyses the NAVE model and explores its capacity to distinguish quality monotonically for three isolated visual distortions: blocking artifacts, Gaussian blur, and white noise. To this end, we create a dataset of short length video sequences containing these three visual artifacts at ten different distortion levels. The main objective is to analyze the NAVE predictions for these sample distortions and verify their consistency. Moreover, by stressing out the model, we expect to discover the limits of the model's performance. Then, we analyze NAVE's predictions in accordance with the different levels of distortions of each isolated visual artifact. We expect to observe a certain coherence between the predictions of the model and the levels of distortion. This can confirm (or deny) if the model is capable of differentiating visual impairments at different distortion levels. In addition, we are able to observe the performances at each range of distortion levels and artifacts, which will allow us to establish certain boundaries for the model. Overall, the findings of this experiment will serve for the improvement of the NAVE model.

The remainder of this document is divided as follows. First, a brief description of the NAVE's metric architecture is presented. Then, the synthetic dataset creation process is described along with the corresponding subjective responses. Next, the NAVE pre-trained model is tested over the synthetic dataset and the corresponding results are presented and discussed. Finally, we list the most relevant conclusions of this study.

NAVE Metric

The No-reference Autoencoder VidEo (NAVE) is a NR-VQA metric. This metric uses a deep autoencoder approach to generate a set of visual descriptive features and estimates the video quality of a video signal using these features. NAVE's architecture, presented in Figure 1, is composed of a two-layer autoencoder module and a mapping function. The metric receives, as input, a set of visual features which are composed by Natural Scene Statistics (NSS) and temporal-spatial measurements. The autoencoder module takes this set of features and generates a new encoded representation with a lower dimension and a higher description capacity. The new set of features is then mapped into video quality scores using a classification function. NAVE takes advantage of the ability of autoencoders in finding relevant properties among input features and producing a stronger set of descriptive features, which are used to produce more accurate video quality predictions.

NAVE was trained over the UnB-AVQ-Experiment1 dataset, which is a large dataset containing audio-visual sequences (video plus accompanying audio) with their corresponding subjective scores [5]. The dataset is composed of 720 processed video se-

Table 1: Specifications of the degradation parameters used to generate the Synthetic Dataset.

		Blocking	**Blurring**		**White Noise**
		Blocking Factor	*kernel*	*deviation*	*sigma*
Test Conditions	TC01	0.3	3	0	0.39
	TC02	0.4	3	400	0.71
	TC03	0.5	5	400	1.07
	TC04	0.6	7	400	1.6
	TC05	0.7	9	400	2.4
	TC06	0.8	11	200	3.61
	TC07	0.9	13	0	5.41
	TC08	1	13	400	8.11
	TC09	1.1	19	0	12.17
	TC10	1.3	19	400	18.25
SRC video	s1	s1	s1		-
	s2	s2	s2		-
	s3	-	s3		s3
	s4	-	s4		s4
	s5	s5	-		s5
	s6	s6	-		s6

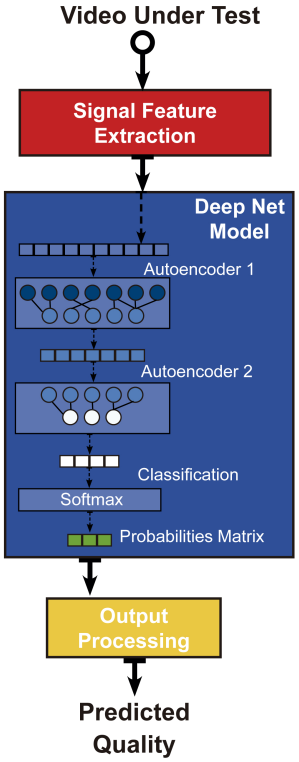


Figure 1: Architecture of the NAVE quality metric.

quences (PVSs) from 60 source sequences (SRCs). The video sequences have a spatial resolution of 720p, a temporal resolution of 30 fps, and a 4:2:0 colour space. The UnB-AVQ-Experiment1 dataset contains video-only combinations of compression and transmission distortions. All video sequences were compressed using H.265 and H.264 codecs at different levels. Packet-loses and frame-freezing effects were added manually to simulate transmission errors. NAVE's performance was tested over two datasets: UnB-AVQ-Experiment1 [5] and LiveNetflix [6]. LiveNetflix dataset is an audio-visual dataset that is composed of 420 PVSs (1080p, 24 fps, 4:2:0) with different transmission conditions and bitrate adaptation strategies. NAVE showed a good performance not only over UnB-AVQ-Experiment1, but also over the external LiveNetflix dataset.

These initial results encouraged researchers to explore the usage of autoencoders to generate stronger sets of features that can be used to produce more accurate quality predictions. A set of ablation experiments showed the capacity of autoencoder-based features to predict not only video but audio-visual quality as well [7]. However, much exploration is needed in order to confirm if NAVE's performance depends on the content of the video sequences or if its performance is in fact related to the visual distortions. In this study, the NAVE metric is tested over a dataset containing both external content and synthetic distortions, the aim is to stress the metric and observe its performance in a completely external context. These findings will serve as a basis to continue studying the usage of autoencoder-based video quality metrics.

Synthetic Database

To study the performance of the NAVE quality model and assess its capacity to distinguish different visual distortions, a synthetic database containing isolated artifacts was created. The Synthetic Database was created using a set of six (6) short length video sequences (8 seconds). The sequences have a spatial resolution of 720p, a temporal resolution of 25 fps, and a 4:2:0 color format. The source videos were collected from the VQEG HDTV Database available at the Consumer Digital Video Library site (www.cdvl.org). These video sequences include three (isolated) types of visual distortions: blocking artifacts, Gaussian blur, and white noise. For each type of distortion, ten different levels of distortions were generated in a synthetic fashion, maintaining a clear distinction between each level. The level of distortion of these visual artifacts increases monotonically with every noticeable distortion level. The Synthetic Database is composed of 120 PVSs (40 Blocking, 40 Blurring, and 40 W. Noise) plus 6 SRCs, adding up to a total of 126 video sequences. Details regarding the parameters of all ten (10) test conditions are presented in Table 1. In addition, sample frames showing the different test condi-

Figure 2: Sample frames of test videos of the Synthetic Database containing Blocking, Blurring, White Noise degradations with different levels of distortion (TC01, TC03, TC05, TC07, TC10).

tions included in the Synthetic Database are depicted in Figure 2. It is worth mentioning that SRC video sequences used to create the Synthetic Database were all different from the SRC video sequences used in the UnB-AVQ-Experiment1 dataset. This was done to prevent using content that was familiar to the (trained) NAVE metric.

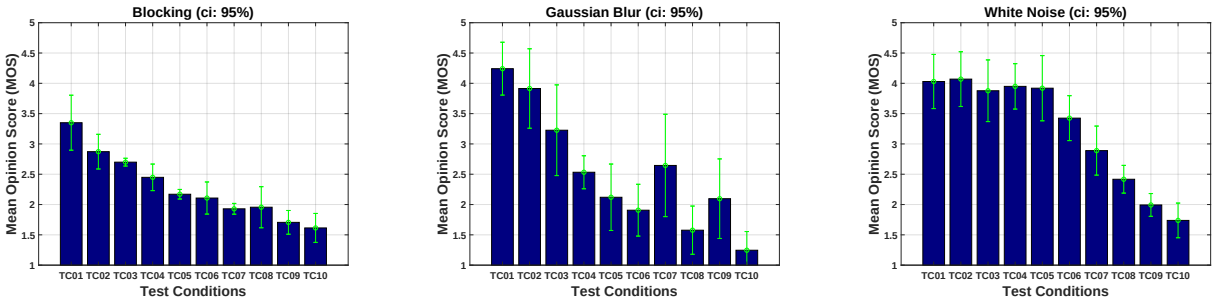
The selection of the levels of distortion was performed empirically by video quality experts through a short subjective test. The experiment was conducted at the University of Brasilia (UnB), in a quiet and isolated room at the Grupo de Processamento de Sinais Digitais (GPDS) of the Department of Engineering (ENE). Hardware equipment consisted of a desktop computer and a LCD monitor. The experiment was conducted with 8 participants, all of them working in video quality assessment area. [1] The experiment was divided in two sessions: the training session and the main experimental session. During the training session participants were presented with examples of the test conditions. This gives participants an idea of the quality range of the PVSs of the experiment and exposes them to the three types of distortion included in the experiment. After each PVS is displayed, the interface shows a quality rating scale to the participants, who are then asked to rate the quality of the PVS using a five-point Absolute Category Rating (ACR) scale ranging from 1 to 5. The rating scale was labeled from 1 to 5 as "Bad", "Poor", "Fair", "Good", and "Excellent". After the training session, the participant is hopefully familiarized with the test methodology and the rating procedure. During the main experimental session, the actual experiment was carried out following the same procedure used in the training session. A single stimulus methodology was used and each sequence was played only once. Sequences were grouped in three sets according to their corresponding distortion: Gaussian Blur, white noise, and blocking. Within each group, sequences were presented randomly to the participants. On average, a single experiment lasted around 40 minutes.

Subjective responses were collected and the associated Mean Opinion Score (MOS) values were computed for each test condition (of all three types of distortion). Figure 3 depicts the MOS results obtained, grouped by type of distortion. Notice from Figure 3(a) that the different test conditions for the blocking distortion were easily distinguished by the participants. In fact, among all other types of distortion, blockiness is the only distortion that shows a strictly monotonic behavior across all test conditions. Similarly, we can notice from Figure 3(b) that MOS values for blurring distortions have a monotonic behavior, although test conditions TC07 and TC09 did not followed this pattern. Finally, as depicted in Figure 3(c), participants seemed to have more trouble distinguishing test conditions with white noise distortions, at least for the first group (TC01 to TC05). Another relevant point to mention is the quality range for different distortion types. Blocking distortions obtained quality scores up to 3.5 points in the MOS scale, meanwhile blurring and white noise distortions obtained quality scores above 4 points in the MOS scale. Overall, quality scores obtained for sequences containing these three visual artifacts has a wide MOS range, which is an important requirement to verify the performance of the pre-trained NAVE metric.

NAVE Performance Analysis

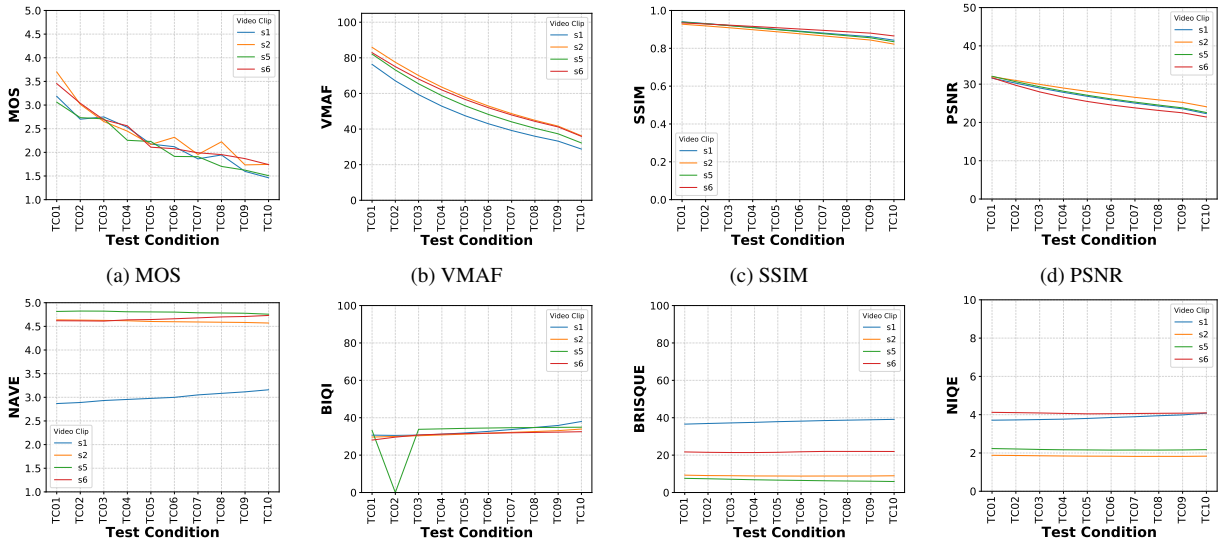
In this section, we present the results of testing the NAVE metric on the Synthetic Database detailed in the previous session. To obtain the quality predictions, we extracted sets of features from the Synthetic Database PVSs and passed them to a pre-trained NAVE metric. We compared NAVE's quality predictions

[1] Due to the current Covid-19 scenario, special care was taken while conducting the experiment. Only one participant was allowed during each experimental session. Surfaces (e.g., keyboard, mouse, desks and chairs) were wiped with disinfectant regularly. Sanitizing hand rub dispensers were made available to all participants. Finally, the room where the experiment was conducted was isolated from other parts of the building and had natural ventilation.



(a) (b) (c)

Figure 3: Mean Opinion Score (MOS), for the different Test Conditions (see Table 1).



(a) MOS (b) VMAF (c) SSIM (d) PSNR

(e) NAVE (f) BIQI (g) BRISQUE (h) NIQE

Figure 4: Synthetic Database - Blocking. Metrics' performance per test condition and SRC video contents.

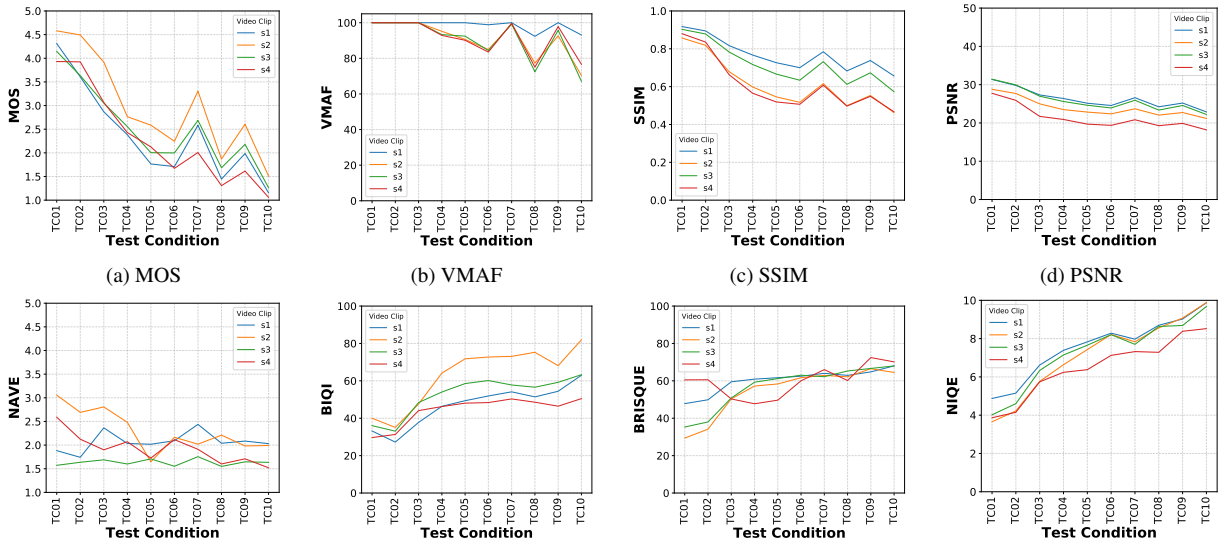
obtained for the Synthetic Database with the predictions obtained with the following set of visual quality metrics:

- FR video metrics: VMAF [8];
- FR image metrics (adapted for video): SSIM [9] and PSNR; and
- NR image metrics (adapted for video): BIQI [10], NIQE [11] and BRISQUE [12].

It is worth mentioning that, similarly to NAVE, all these metrics were designed with different purposes and using different content materials and distortions. Considering that one of the goals of this study is to analyse the behavior of NAVE using 'foreign' contents and distortions, we included a wide variety of visual quality metrics in this analysis, each having different design characteristics.

To understand how the predictions varied across all test conditions, we grouped the metrics' quality predictions by their corresponding SRC video content and presented the results for each distortion separately. First, for the PVSs containing blocking distortions, Figure 4 depicts the MOS versus test condition graph and the corresponding quality predictions versus test condition graphs obtained for NAVE and all tested metrics. Notice from Figure 4(e) that the NAVE metric did not perform very well, not being able to estimate the quality changes the 10 test conditions. A similar performance was observed for the other NR metrics BIQI, BRISQUE and NIQE, as depicted in Figures 4(f), 4(g), and 4(h). On the other hand, as shown in Figure 4 (b), the FR video quality metric VMAF displays a good performance, with quality predictions monotonically decreasing for the 10 test conditions, as expected. This behavior is very similar to the one observed for the MOS responses in Figure 4 (a). As for the other FR quality metrics (SSIM and PSNR), they seem to be able to detect the quality changes, but show a very narrow quality range, as can be observed in Figures 4 (c) and (d).

Figure 5 presents the metrics' quality predictions versus test condition plots for the Gaussian blur distortions. As can be observed in Figure 5(e), although in this case NAVE was able to distinguish the several levels of distortion, the quality predictions values do not decrease monotonically for the 10 test conditions. In Figures 5(f), 5(g), and 5(h), we can notice that the NR image quality metrics BIQI, BRISQUE and NIQE were able to distinguish some distortion levels, presenting more reasonable quality predictions for the test conditions. As for the results obtained for the FR video metric VMAF, which are shown in Figure 5(b)), for the blurring distortions the performance was not as good as for the blockiness distortions. In fact, its range of quality prediction was much narrower. On the other hand, the quality predictions obtained with SSIM and PSNR, shown in Figures 5 (c)



(a) MOS (b) VMAF (c) SSIM (d) PSNR

(e) NAVE (f) BIQI (g) BRISQUE (h) NIQE

Figure 5: Synthetic Database - Gaussian Blur. Metrics' performance per test condition and SRC video contents.

and 5(d), maintained a monotonically decreasing behavior across the 10 test conditions, similar to what was obtained for blockiness distortions.

Finally, Figure 6 presents plots of the metrics' quality predictions versus test conditions for the white noise distortions. Notice from Figure 6(e) that NAVE was able to distinguish the variations across some of the test conditions. However, this was not the case for all SRC video contents. As shown in Figures 6(f), 6(g), and 6(h), the NR metrics BIQI, BRISQUE, and NIQE had a better performance, being able to distinguish the quality levels of almost all test conditions. More specifically, NIQE presented 'well-behaved quality predictions ', showing a monotonically increasing curve across the test conditions. In fact, its results are very similar to the MOS graph displayed in Figure 6(a). Surprisingly, VMAF did not perform very well for white noise distortions, failing to distinguish the different levels of distortions, as shown in Figure 6(b). As for SSIM and PSNR, both metrics had a consistent performance, showing a monotonically decreasing curve across the 10 test conditions, as seen in Figures 6(c) and 6(d).

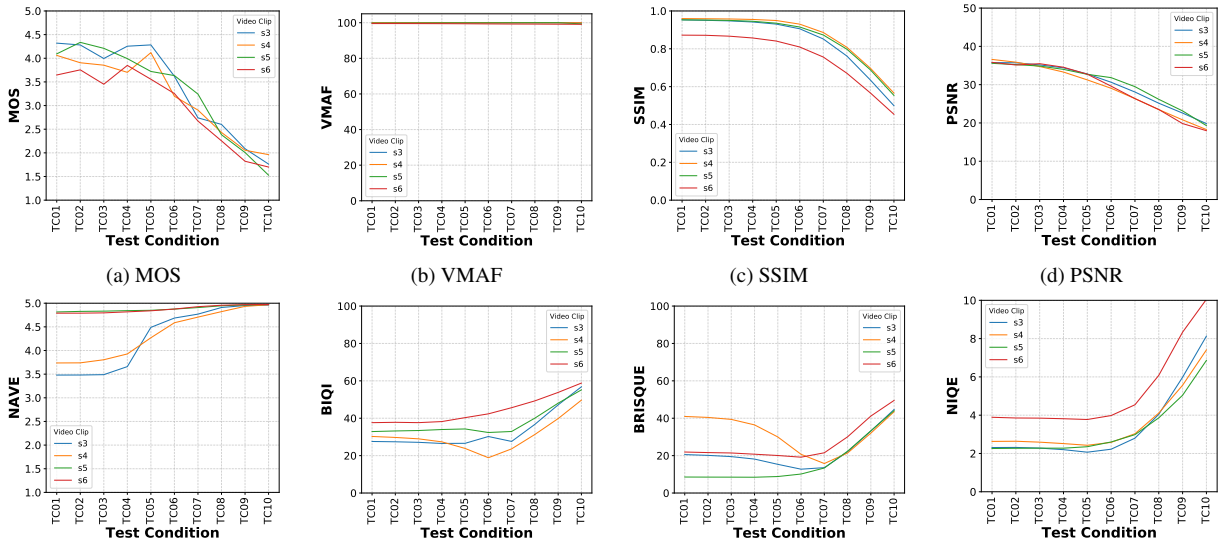
Discussion

Results presented in the previous section exposed a number of points. First, NAVE's limited performance across all three types of distortion is the most notorious one. However, the observed performance of all the other metrics exposed a few more interesting points relating to the objective quality metrics' design and their performance on 'foreign' contents and distortions. We believe a discussion on these issues may bring interesting insights that might be beneficial, not only for the improvement of the NAVE metric but also for the development of more accurate quality models.

NAVE's training was done using a large variety of visual contents. The UnB-AVQ-Experiment1 Database contains 60 different SRC video sequences. Although the intention of training NAVE over such a variety of content was to make the model robust across different types of content, there is still a risk of overfitting the model. The Synthetic Database was created using six SRC sequences takenfrom the VQEG HDTV Database. The content of these SRC sequences was different from the content used during the training of the NAVE metric. This had the goal of guaranteeing that the metric's predictions were solely responding to the visual distortions and not to the content itself. Results presented in Figures 4, 5 and 6 showed that NAVE was able to generalize their predictions across the different PVSs, except for the distortions associated with the Gaussian blur. That is, predictions were coherent across the video content. Further experiments can be performed with an entirely new database, for example using the distortions from the Synthetic Database on the source sequences from the UnB-AVQ-Experiment1. This new database might help to understand the dependency of the NAVE metric on the content used during its training.

As discussed in Section 2, NAVE was trained using three types of visual distortions: packet-loss, frame-freezing, and video compression. Out of these distortions, two of them can be classified as temporal distortions (packet-loss and frame-freezing). This might have influenced on the poor results reported by NAVE for the synthetic distortions, considering that they are all purely spatial degradations. In fact, in a previous study NAVE reported good results when tested on the LiveNetflix Database, which contains several temporal distortions (transmission errors and bitrate adaptation). Re-training the model and including these distortions in an isolated way, might give us a lead on the capacity of NAVE to extend its knowledge domain.

The additional quality metrics used in this work for comparison purposes were the following: VMAF [8], SSIM [9], PSNR, BIQI [10], NIQE [11], and BRISQUE [12]. These metrics were developed to tackle different types of scenarios, distortion, and contents. Some of the metrics were FR metrics (VMAF, SSIM, and PSNR) and others NR metrics (NAVE, BIQI, BRISQUE, and NIQE). Most included some machine learning training (VMAF, NAVE, BIQI, BRISQUE, and NIQE), but we also included simpler metrics that did not require training (SSIM, and PSNR). Our



(a) MOS (b) VMAF (c) SSIM (d) PSNR

(e) NAVE (f) BIQI (g) BRISQUE (h) NIQE

Figure 6: Synthetic Database - White Noise. Metrics' performance per test condition and SRC video content.

goal was to analyze how such a diverse group of quality metrics performed for the Synthetic Database.

Considering the PVSs containing blocking artifacts, we notice that all three FR metrics showed a good performance. Although their predictions varied across all test conditions, the corresponding curves were similar to the MOS curves obtained from the experiment subjective responses. This was not the case for the NR metrics, whose quality predictions were not able to capture the quality variations in the 10 test conditions. This is very interesting since all four NR quality metrics use natural scene statistics (NSS) features as input for their quality predictions. These results might suggest that such measurements have more difficulties to capture quality changes introduced by blocking artifacts.

As for the Gaussian blur distortions, a diverse behavior was observed across all metrics. From the FR metrics, SSIM and PSNR reported predictions varying across all test conditions, while VMAF had some trouble distinguishing the different quality levels. For this type of distortion, NR metrics performed much better than for blocking artifacts. BIQI, BRISQUE, and NIQE were capable of recognizing the different levels of distortion. This aligns with the good performance reported over the Live IQA Database [10, 11, 12].

For the white noise distortions, the adapted image metrics (PSNR, SSIM, BIQI, BRISQUE, and NIQE) showed a far better performance than the video quality metrics (VMAF and NAVE). As with the previous distortions, the quality predictions generated by the FR metrics SSIM and PSNR decreased monotonically with the test conditions, as expected . The same behavior was also observed for the NR metrics BIQI, BRISQUE, and, NIQE. This performance should not be surprising, considering that the design of image quality metrics often include white noise distortions, which is a common degradation in most image quality database such as LIVE [13] and TID [14]. This might also explain the poor performance of VMAF for sequences with white noise distortions. VMAF is a machine learning based quality metric designed for video streaming application, which in most cases do not have white noise distortions. It is worth pointing out that noise is sometimes used by the film industry for artistic purposes, which do not affect the quality of the experience. Therefore, quality metrics used for these applications might less sensitive to noise.

Conclusions

This study aimed to test the performance of NAVE, a NR video quality metric. In order to do so, the Synthetic Database was created, containing distortions and SRC video contents completely different from the ones used in NAVE's training. Such distortions included blocking artifacts, Gaussian blur, and white noise, with ten different test conditions of increasing impairment levels. A subjective experiment was performed to collect subjective quality scores for the dataset. Results showed that NAVE had difficulties distinguishing quality variations across all three types of distortions. These results bring up some interesting questions that require further exploration. For example, a possible future work would be recreating the test conditions of the Synthetic Database on the SRC video content that was used to train NAVE. Also, it would be interesting to check the performance of trained-based metrics on different distortions at different levels, including those for which the metrics were trained on. We believe this type of study is still lacking in the literature.

Acknowledgments

This work was supported in part by the Conselho Nacional de Desenvolvimento Cientfico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the University of Brasília (UnB), the research grant from Science Foundation Ireland (SFI) and the European Regional Development Fund under Grant Number 12/RC/2289_P2 and Grant Number 13/RC/2077_P2.

References

[1] P. Gastaldo and J. A. Redi, "Machine learning solutions for objective visual quality assessment," in *6th international workshop on video*

processing and quality metrics for consumer electronics, VPQM, vol. 12, 2012.

[2] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[3] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[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)*. IEEE, 2019, pp. 1755–1759.

[5] H. B. Martinez, A. Hines, and M. C. Farias, "Unb-av: An audio-visual database for multimedia quality research," *IEEE Access*, vol. 8, pp. 56 641–56 649, 2020.

[6] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards perceptually optimized end-to-end adaptive video streaming," *arXiv preprint arXiv:1808.03898*, 2018.

[7] H. Martinez, A. Hines, and M. C. Q. Farias, "How deep is your encoder: An analysis of features descriptors for an autoencoder-based audio-visual quality metric," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[8] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, p. 2, 2016.

[9] 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, 2004.

[10] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal processing letters*, vol. 17, no. 5, pp. 513–516, 2010.

[11] 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.

[12] 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.

[13] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[14] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015.