Analyzing and Exploring Patterns on Extracted Features used on a Quality Assessment Model

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

[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 features used as input of this ML-based model looking for possible patterns discernible 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 performed an analysis of the extracted features using a 2-D visualization format derived from the feature matrices. This procedure allowed us to detect the different patterns generated by the types of features and visual distortions, identifying the most important features for quality estimation.

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 how the visual features used in NAVE respond to visual impairments and the content itself.

To help answer some of these questions, this paper carries out an analysis of the visual features used in NAVE's architecture using a set of 2-D representations of the extracted features. To this end, we create a dataset of short length video sequences containing three visual artifacts: blocking artifacts, Gaussian blur, and white noise. The paper analyses the extracted features and explores its capacity to distinguish these isolated visual distortions at ten different quality levels.

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 features are tested over the synthetic dataset and the corresponding visualizations 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 sequences (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

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

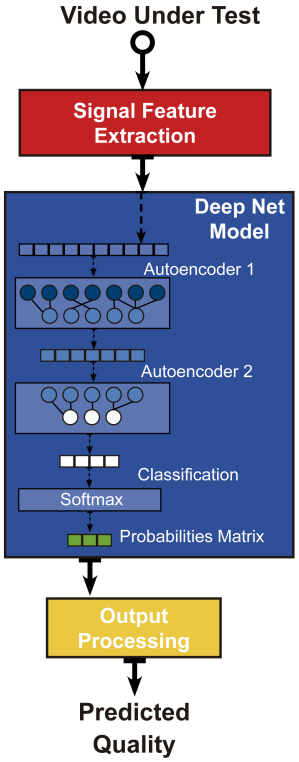


Figure 1: Architecture of the NAVE quality metric.

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.

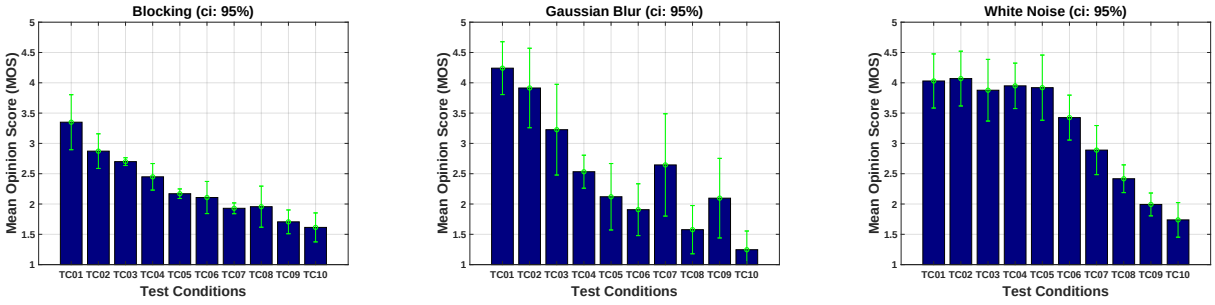
These initial results encouraged researchers to explore the usage of autoencoders to generate stronger sets of features that can be used to produce more accurate quality predictions. A set of ablation experiments showed the capacity of autoencoder-based features to predict not only video but audio-visual 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 visual input features used in the NAVE metric are analysed over a dataset containing synthetic distortions. The aim is to check how these features respond to different types of distortions to help identify which of these features can be used to estimate quality. These findings will serve as a basis to continue studying the usage of autoencoder-based video quality metrics.

Synthetic Database

To study the capacity of the NAVE quality model 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 conditions included in the Synthetic Database are depicted in Figure 2.

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

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).



(a) (b) (c)

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

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

[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.

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.

Feature Analysis

The NAVE model uses 88 visual features as input extracted using the Diivine algorithm [15]. Each one of them serves to represent relevant visual characteristics of the video sequence, so it is expected that they respond to the presence of visual distortions. Table 2 shows a brief description and the corresponding computation procedure for all the extracted visual features. Features were organized in six groups, according to their main characteristics. In this section, it will be presented the results of aggregating and analyzing the behaviour of the features, clustering by type of distortion and content of the video.

For visualization purposes, a pre-processing step was necessary to normalize the feature values so they can be presented in a single image. First, the minimum value was calculated and the total range of a specific feature, i.e., given that the group of features is a 88×20 dimension matrix, we normalized a 20-position vector based on every possible value of that specific feature. For this, the minimum value was added to every position of the array, so it would shift all the values to a new minimum of zero. The next step was to get the total range of that specific feature, calculated by the maximum value minus the minimum value. So, with the range in hands, it was possible to normalize all the data so it would be not only non-negative numbers, but within a range of zero and one. Given that we now have all the features normalized by their own group of values, we just multiply them by 255 and now we finally have features that can be visualized as images in a way that is normalized by each group, then the comparison can be made between one and the other.

In Figures 4, 5 and 6 it is possible to observe that all of the features show different patterns, depending on what degradation was applied to the source video. Not only that, but it's also possible to analyse that each group of features is well represented and presents very distinguishable pattern from each other. So, even with the source video's feature (first figure from left to right), it is very clear the edges of each feature group. In order to study further the behavior of features with the degradations, it was also plotted, in Figures 7, 8 and 9, the difference between the source original video's feature and every other degraded feature. This is way it is easier to visualize the movement of features as the video becomes more and more degraded.

Results Discussion

It is possible to observe that some group of features are more affected with certain kinds of distortion. For instance, we can analyse that specifically Block Distortion does not have as much effect on features as other types of degradation, such as Blur and Noise. As we can see from the difference representations from the original value in Figures 7, 8 and 9. Moreover, we can observe that some specific groups of features are more or less affected depending on the type and intensity of distortion. For instance, it was quite clear which groups were more affected by blocking artifacts represented in Figure 7. Both groups f_{13}-f_{24} (shape parameter of subband coefficients) and f_{44}-f_{73} (spatial correlation across subbands) were gradually affected by blocking effect, which means that shape parameter of subband coefficients and spatial correlation across subbands have greater sensitivity to this kind of degradation.

On the other hand, almost every group of features were affected by Blur distortion. It is possible to draw attention to the groups of shape parameter of subband coefficients and spatial correlation across subbands, as these were the most affected even at low levels of degradation. As the intensity of the degradation increases, another group of features stands out, f_1-f_{12} (variance of subband coefficients) starting to behave quite differently from the original value. Analysing Figure 9, the differences from the original value for Noise distortion for features labeled as shape parameter of subband coefficients and spatial correlation across subbands were the most affected at low levels of distortion. As the intensity increases, the f_{32}-f_{43} group (correlation across scales) stands out, presenting increasingly higher values of difference from the original each time. Finally, as the sigma value (Noise factor) increases, the spatial correlation across subbands start to line up, as if each feature vector takes on a single value. That behavior can be further studied in the future, as the general behavior of the features represented by images, as presented in this study.

Conclusions

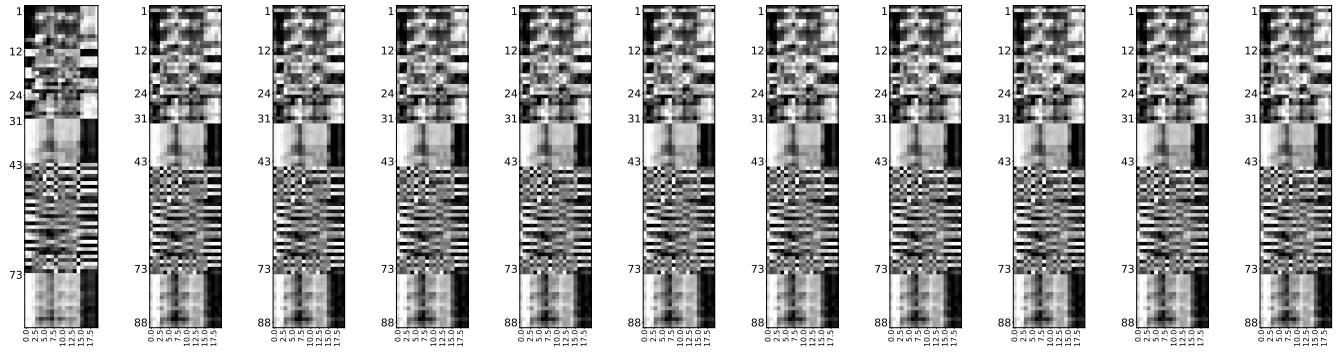
This study aimed to test the behavior of extracted features over different types of degradation applied to a synthetic video database. As presented previously, it was possible to observe that shape parameter of subband coefficients and spatial correlation across subbands groups of features have greater sensitivity to all three types of distortion tested. The reason for that and what are the consequences will be studied in the future. Another important result was that the correlation across subbands is heavily influenced by Noise distortion, but only from high sigma values. And, correspondingly, the variance of subband coefficients stands out when distorted with Blur Degradation. In contrast, with the exception of the shape parameter of subband coefficients and spatial correlation across subbands groups, that are the two groups of features affected by all distortions, Block distortion had almost no impact on the features when compared to the original video.

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.

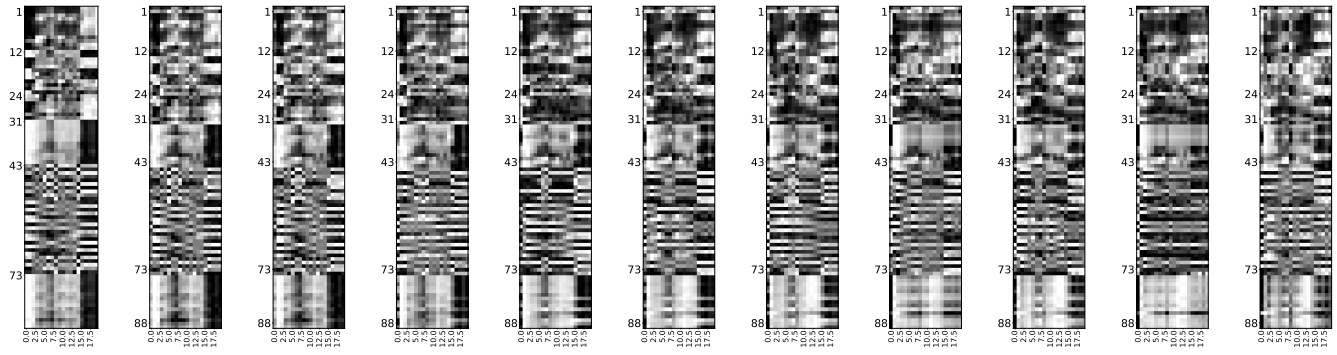
Table 2: Specifications of the features generated from the Synthetic Dataset.

Feature ID	Description	Computation Procedure
f_1-f_{12}	Variance of subband coefficients	Fitting a generalized Gaussian to subband coefficients.
f_{13}-f_{24}	Shape parameter of subband coefficients	Fitting a generalized Gaussian to subband coefficients
f_{25}-f_{31}	Shape parameter across subband coefficients	Fitting a generalized Gaussian to orientation subband coefficients.
f_{32}-f_{43}	Correlations across scales	Computing windowed structural correlation between filter responses.
f_{44}-f_{73}	Spatial correlation across subbands	Fitting a polynomial to the correlation function.
f_{74}-f_{88}	Across orientation statistics	Computing windowed structural correlation between adjacent orientations at same scale.



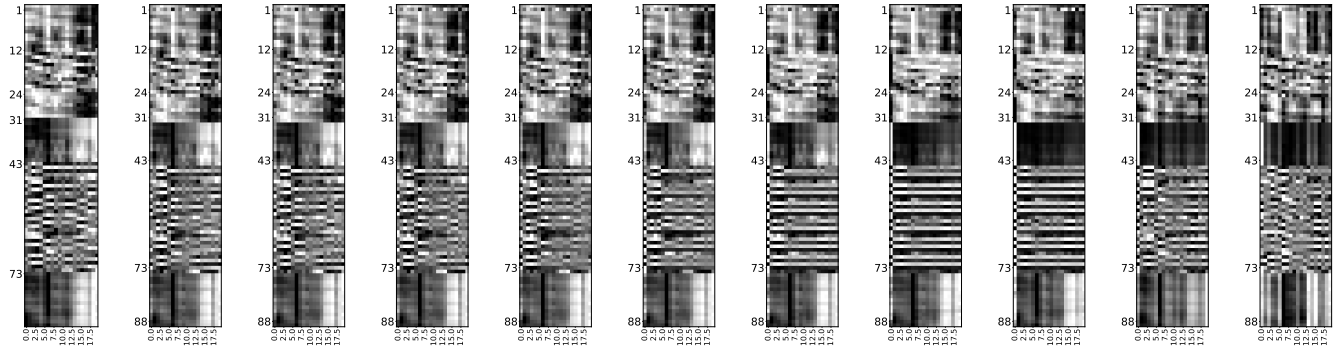
(original) (0.3) (0.4) (0.5) (0.6) (0.7) (0.8) (0.9) (1) (1.1) (1.3)

Figure 4: 2D feature representation from Block distortion videos derived from a 88-by-20 matrix where 88 represent the number of features and 20 corresponds to the mean sampled frames of the original video. Source video to most degraded video from left to right.



(original) (3,0) (3,400) (5,400) (7,400) (9,400) (11,200) (13,0) (13,400) (19,0) (19,400)

Figure 5: 2D feature representation from Blur distortion videos.



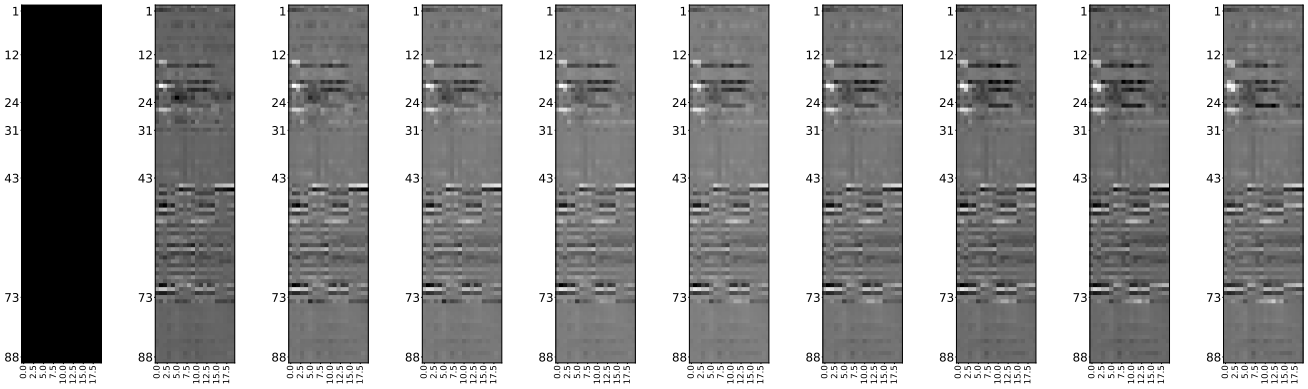
(original) (0.39) (0.71) (1.07) (1.6) (2.4) (3.61) (5.41) (8.11) (12.17) (18.25)

Figure 6: 2D feature representation from Noise distortion videos.

References

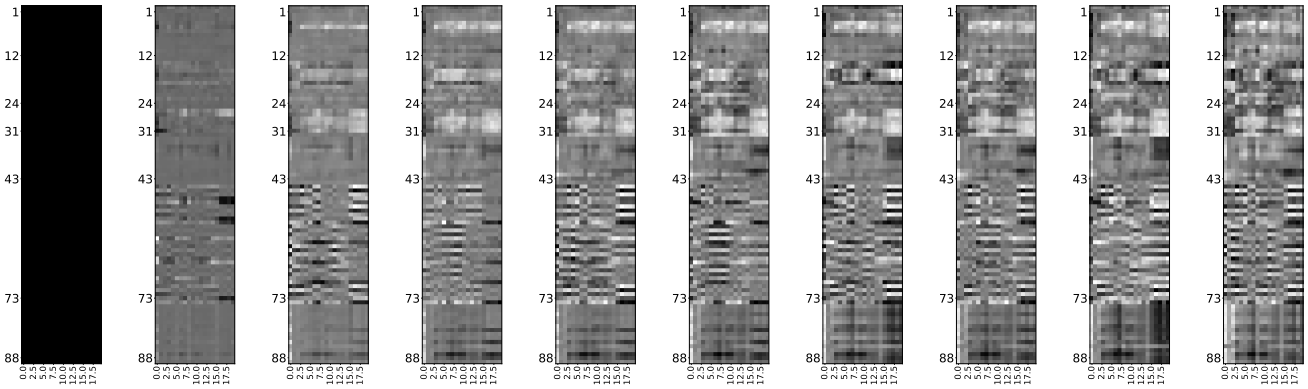
[1] P. Gastaldo and J. A. Redi, "Machine learning solutions for objective visual quality assessment," in *6th international workshop on video processing and quality metrics for consumer electronics, VPQM*, vol. 12, 2012.

[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.



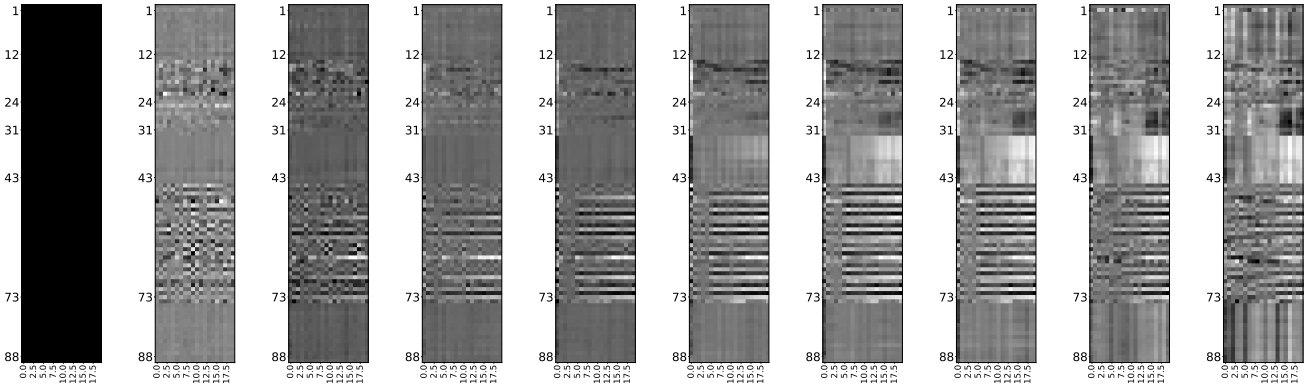
(0.3) (0.4) (0.5) (0.6) (0.7) (0.8) (0.9) (1) (1.1) (1.3)

Figure 7: 2D feature representation from difference with least distorted Video with Increasing Degradation of Block Distortion



(3,0) (3,400) (5,400) (7,400) (9,400) (11,200) (13,0) (13,400) (19,0) (19,400)

Figure 8: 2D feature representation from difference with least distorted Video with Increasing Degradation of Blur Distortion



(0.39) (0.71) (1.07) (1.6) (2.4) (3.61) (5.41) (8.11) (12.17) (18.25)

Figure 9: 2D feature representation from difference with least distorted Video with Increasing Degradation of Noise Distortion

[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, "Im-

age 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.

[15] Yi Zhang, Anush K Moorthy, Damon M Chandler, and Alan 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.