electronics

MDPI

Article

No-Reference Video Quality Assessment Based on Benford's Law and Perceptual Features

Domonkos Varga

Ronin Institute, Montclair, NJ 07043, USA; domonkos.varga@ronininstitute.org

Abstract: No-reference video quality assessment (NR-VQA) has piqued the scientific community's interest throughout the last few decades, owing to its importance in human-centered interfaces. The goal of NR-VQA is to predict the perceptual quality of digital videos without any information about their distortion-free counterparts. Over the past few decades, NR-VQA has become a very popular research topic due to the spread of multimedia content and video databases. For successful video quality evaluation, creating an effective video representation from the original video is a crucial step. In this paper, we propose a powerful feature vector for NR-VQA inspired by Benford's law. Specifically, it is demonstrated that first-digit distributions extracted from different transform domains of the video volume data are quality-aware features and can be effectively mapped onto perceptual quality scores. Extensive experiments were carried out on two large, authentically distorted VQA benchmark databases.

Keywords: no-reference video quality assessment; Benford's law; feature extraction

Citation: Varga, D. No-Reference Video Quality Assessment Based on Benford's Law and Perceptual Features. *Electronics* **2021**, *10*, 2768. https://doi.org/10.3390/electronics10222768

Academic Editor: Krzysztof Okarma

Received: 26 October 2021
Accepted: 10 November 2021
Published: 12 November 2021

1. Introduction

As digital media takes a more central part in our daily lives, research on video quality assessment (VQA) becomes more and more important. For instance, about 70% of the overall Internet bandwidth is occupied by digital video streaming [1]. Moreover, it is predicted that the occupied bandwidth will increase to between 80% and 90% by 2022 [2]. As a consequence, the precise estimation of video quality is of vital importance for video streaming and sharing. In addition, VQA is also crucial in video restoration, reproduction, enhancement, and compression. Hence, the scientific community has devoted much attention and effort to this research field, continuously developing and devising algorithms, methods, and metrics that are able to estimate digital videos' perceptual quality.

The most accurate way to assess video quality is to collect subjective opinions from human observers in a laboratory environment involving experts. This process is called subjective video quality assessment, which is rather time consuming and expensive. However, the scores collected in subjective user studies can be applied as ground-truth data for objective video quality assessment. Namely, objective video quality assessment deals with the construction of algorithms that accurately estimate the perceptual quality of a given video sequence. Depending on the availability of the reference, distortion-free videos, objective VQA methods can be grouped into three classes: full-reference (FR), reduced-reference (RR), and no-reference (NR) ones. As the names indicate, FR-VQA algorithms have full information about the reference videos, while NR-VQA ones do not have access to the reference videos. Moreover, RR-VQA methods can be considered as a transition between FR-VQA and NR-VQA algorithms. Namely, they have partial information about the reference videos, for example in the form of sets of extracted features.

In this paper, we propose a novel NR-VQA algorithm utilizing Benford's law. Frank Benford, a physicist with General Electric, collected over 20,000 numbers in 1938 from extremely various sources, such as Readers' Digest articles, atomic weights, population sizes, drainage rates of rivers, and physical constants [3,4]. It was demonstrated that

the distribution of the first digits follows an algorithmic rule. Moreover, the observation works on a distribution of numbers if that distribution spans over a few orders of magnitude. Benford's law states that the leading d ($d \in \{1, ..., 9\}$) in a natural dataset occurs with probability:

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}\left(1 + \frac{1}{d}\right). \quad (1)$$

Figure 1 depicts the distribution of leading digits in natural datasets predicted by Benford's law. Benford's law has been proven as an efficient tool in digital analysis technology [5,6]. Namely, the first-digit distribution (FDD) in accounting data of companies and unmanipulated macroeconomic data are expected to follow Benford's law [7–9]. First, Jolion [10] investigated Benford's law in the context of digital images. The FDD pixel values do not follow Benford's law, since pixel values are in the interval of $[0, 255]$ and do not span over a few magnitudes. However, the FDD of gradient magnitudes matches well with Benford's law prediction. Similarly, Perez-Gonzalez et al. [11,12] pointed out that the discrete cosine transform (DCT) coefficients of a digital image follow Benford's law. Fu et al. [13] demonstrated that the FDD of DCT coefficients can be used for the detection of distorted JPEG images. Similarly, Andriotis et al. [14] utilized the FDD of DCT coefficients, but they used it for image steganalysis to ascertain whether a digital image contains a hidden message or not. In our previous study [15], we demonstrated that FDDs extracted from different domains (wavelet, DCT, shearlet, singular values) are quality-aware features and they can be used for no-reference image quality assessment.

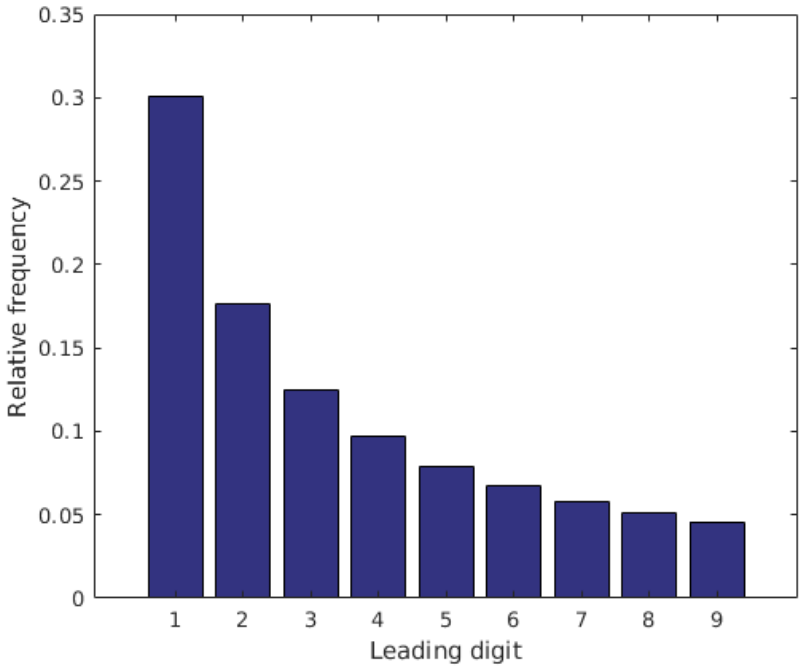


Figure 1. Illustration of Benford's law.

1.1. Contributions

The main contributions of this paper are the following. In our previous works [15,16], the usage of FDD feature vectors was thoroughly investigated in the context of no-reference image quality assessment. Specifically, it was proven that FDDs of different 2D transform domains are quality-aware features and can be mapped effectively onto image quality scores. In contrast, we explored the applicability of video-level FDD-based feature vectors for NR-VQA in this study. We demonstrate that quality-aware FDD feature vectors can be extracted from the video volume data considering different 3D transform do-

mains (spatial, wavelet, discrete cosine transform, discrete Fourier transform, higher-order singular-values). It is demonstrated that fusing FDD-based and perceptual features results in a powerful video representation for NR-VQA.

1.2. Structure

The rest of the paper is organized as follows. An overview is provided about NR-VQA in Section 2. Subsequently, Section 3 contains a detailed description of the proposed method based on Benford's law. Section 4 presents the experimental results and analysis on publicly available VQA databases. Finally, this study is concluded in Section 5.

2. Related Work

As already mentioned, VQA can be classified into two groups: subjective and objective. Specifically, subjective VQA deals with measuring perceptual video quality involving human observers either in a laboratory environment or in a crowdsourcing experiment [17,18], while objective VQA attempts to devise mathematical algorithms that are able to estimate video signals' human perceptual quality. In the literature, one can find many recommendations for choosing video sequences, system settings, and test methodologies for subjective VQA. These are defined in the documents of the International Telecommunication Union, such as ITU-R BT.500 [19], ITU-T P.913 [20], or ITU-T P.910 [21]. The average judgment of the human observers is expressed as the mean opinion score (MOS), which can range from 1–5 or 0–100, for example. Moreover, a higher MOS indicates higher visual quality. As a result of subjective VQA experiments, many VQA databases are publicly available for researchers, such as KoNViD-1k [22] or the LIVE Video Quality Challenge (VQC) [23] database. An overview of publicly available benchmark VQA databases can be found in [24–26].

In the literature, objective NR-VQA algorithms are classified into bitstream-based, pixel-based, and hybrid models [27]. Bitstream-based algorithms extract features directly from encoded video sequences to predict perceptual video quality. For example, Argyropoulos et al. [28] presented an algorithm for the prediction of continuous estimates of the visibility of packet losses. Moreover, the authors demonstrated that visible losses have a significant impact on video quality degradation. Keimel et al. [29] devised a H.264/AVC bitstream algorithm. Namely, multiple features were extracted from the encoded video sequences, such as slice type, bits per slice, average quantization parameter per slice, average, minimum and maximum motion vector length per slice, and average and maximum motion vector error per slice. These features were mapped onto perceptual quality scores using partial least squares regression. Similarly, Chen and Wu [30] identified a set of features, such as concealment error, motion vector concealment error, pixel-level transmission distortion, etc., and conclude from them the perceptual video quality. The approach of Pandremmenou et al. [31] is also similar to the method of Argyropoulos et al. [28], but they defined a wider range of features. Specifically, a feature vector of length 46 was mapped onto perceptual quality scores using the least absolute shrinkage and selection operator regression. In contrast to bitstream-based methods, pixel-based algorithms utilize the raw video signal solely as the input. For example, Saad et al. [32] extracted features using the 3D-DCT. Then, a linear kernel support vector regressor (SVR) was trained to predict the visual quality of videos. Similarly, Zhu et al. [33] presented a DCT-based NR-VQA model. Specifically, frame-level DCT coefficient-based features (peakiness, smoothness, sharpness, etc.) were temporally pooled to compile video-level feature vectors. Next, the video-level feature vectors were mapped onto perceptual quality scores using a shallow neural network. In contrast, Dendi et al. [34] constructed a video representation using the mean subtracted and contrast normalized (MSCN) coefficients of certain spatiotemporal statistics of a natural video. More specifically, an asymmetric generalized Gaussian distribution was fit onto the MSCN coefficients of a natural video and its Gabor bandpass filtered counterpart. The parameters of asymmetric generalized Gaussian distributions were considered as quality-aware features and mapped onto perceptual quality scores

with the help of an SVR. Ebenezer et al. [35] proposed a video representation that contains information about both the spatial and temporal information. Namely, the authors defined space–time chips as quality-aware features, which are cuts from the original video data in specific directions obtained from local motion flow. Subsequently, asymmetric generalized Gaussian distributions were fit to the bandpass histograms of space–time chips. Finally, the parameters of asymmetric generalized Gaussian distributions were mapped onto perceptual quality scores by a trained SVR.

Recently, deep-learning techniques have been proven very successful in image-based object detection [36], semantic segmentation [37], image generation [38], etc. Researchers also have introduced deep-learning-based NR-VQA methods. For instance, Li et al. [39] extracted quality-aware visual features from video blocks using a 3D shearlet transform. These features were fed into convolutional neural networks to predict perceptual video quality. Ahn and Lee [40] combined deep and hand-crafted features for NR-VQA. Specifically, a pretrained convolutional neural network was applied to extract features from video frames, while temporal features were modeled by hand-crafted features. Li et al. [41] introduced a mixed dataset training strategy. Namely, the authors' network was trained on mixed data and addressed by two different loss functions, i.e., monotonicity-induced and linearity-induced losses.

3. Proposed Method

The high-level overview of the introduced NR-VQA algorithm is depicted in Figure 2. As can be seen from this figure, video-level feature vectors are extracted from the training video sequences to train a machine-learning model, which is later utilized in the testing stage to estimate the perceptual quality of unseen videos. In this paper, a novel set of FDD and perceptual quality-aware features are presented for NR-VQA. Section 3.1 describes the proposed FDD-based features, while the applied perceptual features are described in Section 3.2.

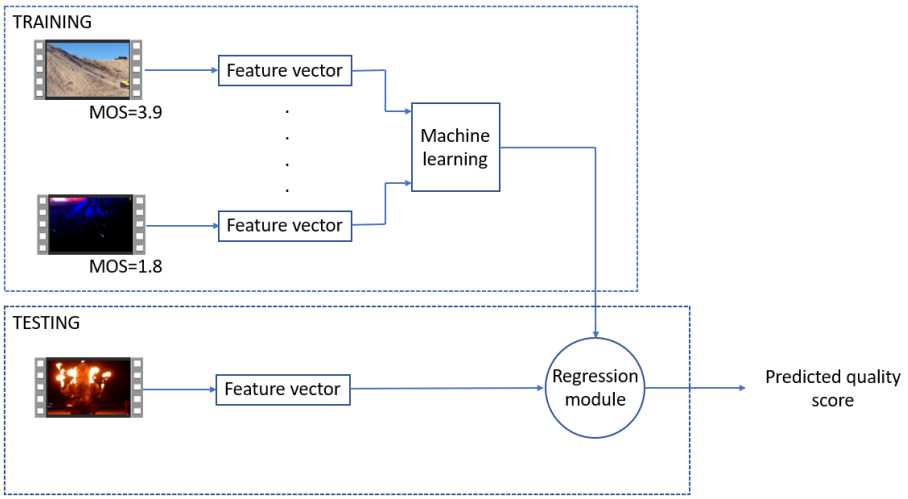


Figure 2. High-level overview of the proposed method. Video-level feature vectors are extracted from the training videos to train a machine-learning model, which is later utilized in the testing phase to predict the quality of previously unseen videos. In this study, we propose the fusion of the Benford-law-inspired first-digit distribution and perceptual features.

3.1. FDD-Based Features

FDD features are extracted from the spatial, 3D wavelet, 3D Fourier, and 3D discrete cosine transform domains of a video sequence. Moreover, the FDD of higher-order singular-values of a video sequence is also considered as a quality-aware feature.

FDDs are extracted in the spatial domain of video sequences using 3D directional gradients. To find the directional gradients of a 3D grayscale video sequence of size $N_1 \times N_2 \times N_3$, the 3D extension of the Sobel edge detector [42] (illustrated in Figure 3) was

applied in this paper. Moreover, the size of the applied 3D Sobel operator for each direction was $3 \times 3 \times 3$. The convolution operators in the x, y, and z directions are given as:

$$S_x(:,:,-1) = \begin{pmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -1 & 0 & 1 \end{pmatrix}, \; S_x(:,:,0) = \begin{pmatrix} -3 & 0 & 3 \\ -6 & 0 & 6 \\ -3 & 0 & 3 \end{pmatrix}, \; S_x(:,:,1) = \begin{pmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -1 & 0 & 1 \end{pmatrix}, \quad (2)$$

$$S_y(:,:,-1) = \begin{pmatrix} -1 & -3 & -1 \\ 0 & 0 & 0 \\ 1 & 3 & 1 \end{pmatrix}, \; S_y(:,:,0) = \begin{pmatrix} -3 & -6 & -3 \\ 0 & 0 & 0 \\ 3 & 6 & 3 \end{pmatrix}, \; S_y(:,:,1) = \begin{pmatrix} -1 & -3 & -1 \\ 0 & 0 & 0 \\ 1 & 3 & 1 \end{pmatrix}, \quad (3)$$

$$S_z(:,:,-1) = \begin{pmatrix} -1 & -3 & -1 \\ -3 & -6 & -3 \\ -1 & -3 & -1 \end{pmatrix}, \; S_z(:,:,0) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; S_z(:,:,1) = \begin{pmatrix} 1 & 3 & 1 \\ 3 & 6 & 3 \\ 1 & 3 & 1 \end{pmatrix}. \quad (4)$$

If we define $\mathbf{V}$ as the grayscale video sequence of size $N_1 \times N_2 \times N_3$, and $\mathbf{G}_x$, $\mathbf{G}_y$, and $\mathbf{G}_z$ are three arrays, which at each point consist of the x, y, and z directional derivative approximations, the computations are as follows:

$$\mathbf{G}_x = \mathbf{S}_x * \mathbf{V}, \; \mathbf{G}_y = \mathbf{S}_y * \mathbf{V}, \text{ and } \mathbf{G}_z = \mathbf{S}_z * \mathbf{V}, \quad (5)$$

where $*$ stands for the 3D convolution operator. In Figure 3, the 3D Sobel operators and filtering are depicted where the x, y, and z axes correspond to the crossline, inline, and time axes of a grayscale video sequence, respectively. In the spatial domain, three FDDs are obtained from $\mathbf{G}_x$, $\mathbf{G}_y$, and $\mathbf{G}_z$.

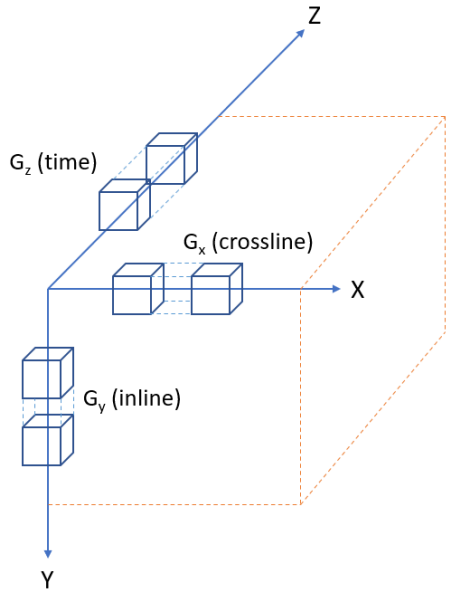


Figure 3. Illustration of 3D Sobel filtering. The x, y, and z axes correspond to the crossline, inline, and time axes of a grayscale video sequence, respectively.

Discrete wavelet transform (DWT) was devised to correct the resolution problems of the short-time discrete Fourier transform [43]. It has a huge number of applications in engineering and computer science, since it is able to represent signals in a redundant form. The 3D DWT is depicted in Figure 4. As one can see from this figure, the volume data (in our case, a grayscale video sequence) are decomposed into eight sub-bands in the case of single-level processing. These can be grouped into three distinct classes: an approximation sub-band (LLL), spectral variation sub-bands (LLH, LHH, HLH), and spatial variation

sub-bands (LHL, HLL, HHL). Moreover, Daubechies mother wavelets were utilized in our implementation. Quality-aware FDD features were extracted from the spectral variation and the spatial variation sub-bands.

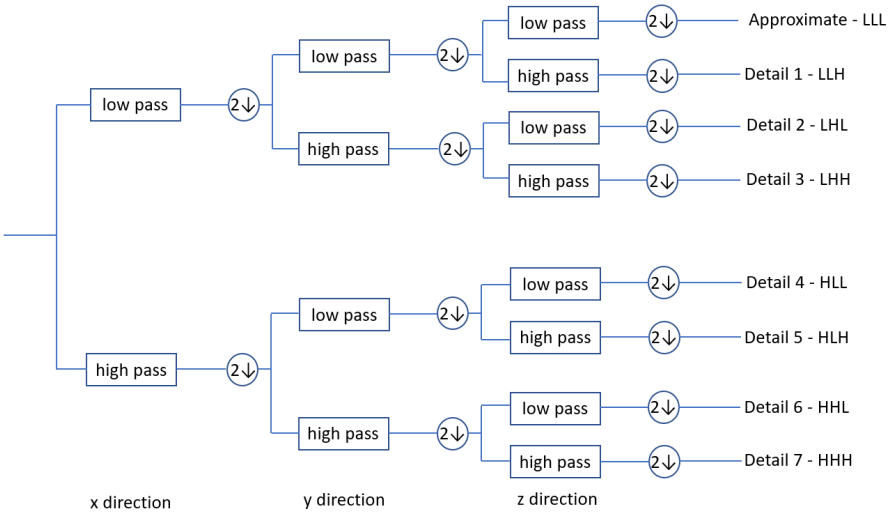


Figure 4. Illustration of the 3D DWT.

The 3D DCT is an extension of the DCT to the three-dimensional space and defined as:

$$X_{k_1,k_2,k_3} = \sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1}\sum_{n_3=0}^{N_3-1} V_{n_1,n_2,n_3} \cos\left[\frac{\pi}{N_1}\left(n_1+\frac{1}{2}\right)k_1\right] \cos\left[\frac{\pi}{N_2}\left(n_2+\frac{1}{2}\right)k_2\right] \cos\left[\frac{\pi}{N_3}\left(n_3+\frac{1}{2}\right)k_3\right], \text{ for } k_i = 0,1,2,...,N_{i-1},\ i = 1,2,3 \tag{6}$$

where X_{k_1,k_2,k_3} are the 3D DCT coefficients and V_{n_1,n_2,n_3} denote the n_1th, n_2th, and n_3th pixel value of the grayscale video sequence $\mathbf{V}$. The FDDs of 3D DCT coefficients are considered as quality-aware features.

Besides 3D DCT coefficients, 3D discrete Fourier transform (DFT) coefficients of grayscale video sequences were also used to extract an FDD feature vector for video representation. The DFT of a grayscale video sequence $\mathbf{V}$ (three-dimensional array of size $N_1 \times N_2 \times N_3$) is defined as:

$$Y_{k_1,k_2,k_3} = \sum_{n_1=0}^{N_1-1} \omega_{m_1}^{k_1,n_1} \sum_{n_2=0}^{N_2-1} \omega_{m_2}^{k_2,n_2} \sum_{n_3=0}^{N_3-1} \omega_{m_3}^{k_3,n_3} V_{n_1,n_2,n_3}, \tag{7}$$

where Y_{k_1,k_2,k_3} stand for the DFT coefficients, V_{n_1,n_2,n_3} denote the n_1th, n_2th, and n_3th pixel value of the grayscale video sequence $\mathbf{V}$, $\omega_{m_k} = e^{-\frac{2\pi i}{m_k}}$ $(m_k = 1,2,3)$ are the complex roots of unity, and i is the imaginary unit. In our implementation, DFT coefficients were determined by the fast Fourier transform [44]. In our study, the FFD of DFT coefficients were used as quality-aware features.

Higher-order singular-value decomposition (HOSVD) can be considered as one generalization of matrix singular-value decomposition [45]. Namely, a tensor's HOSVD corresponds to a specific orthogonal Tucker decomposition [46,47]. Every tensor $\mathbf{V}$ of size $N_1 \times N_2 \times N_3$ can be written as:

$$\mathbf{V} = \sum_{n_1=1}^{N_1}\sum_{n_2=1}^{N_2}\sum_{n_3=1}^{N_3} \sigma_{n_1,n_2,n_3}(\mathbf{u}_{n_1}^{(1)} \circ \mathbf{u}_{n_2}^{(2)} \circ \mathbf{u}_{n_3}^{(3)}), \tag{8}$$

where ∘ stands for the outer product operation and:

$$\mathbf{U}^{(1)} = [\mathbf{u}_1^{(1)}, \mathbf{u}_2^{(1)}, ..., \mathbf{u}_{N_1}^{(1)}], \tag{9}$$

$$\mathbf{U}^{(2)} = [\mathbf{u}_1^{(2)}, \mathbf{u}_2^{(2)}, ..., \mathbf{u}_{N_2}^{(2)}], \tag{10}$$

$$\mathbf{U}^{(3)} = [\mathbf{u}_1^{(3)}, \mathbf{u}_2^{(3)}, ..., \mathbf{u}_{N_3}^{(3)}], \tag{11}$$

are three unitary matrices of sizes $N_1 \times N_1$, $N_2 \times N_2$, $N_3 \times N_3$, respectively. The FDD of singular values (σ_{n_1,n_2,n_3}'s) was used as the quality-aware feature.

In Tables 1–7, the mean FDDs of X directional gradient magnitudes, Y directional gradient magnitudes, HLL wavelet coefficients, HHL wavelet coefficients, 3D DFT coefficients, 3D DCT coefficients, and higher-order singular-values are summarized with respect to five equal MOS intervals of KoNViD-1k [22]. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality. It can be observed that the FDDs of videos with a higher quality fit better the prediction of Benford's law. Moreover, the distance between the actual FDD and the Benford law prediction is also lower in the case of high-quality videos. The distance is given using the symmetric Kullback–Leibler (sKL) divergence.

Table 1. Mean FDD of X directional gradient magnitudes in KoNViD-1k [22] with respect to different MOS intervals. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality. In the last column, the symmetric Kullback–Leibler (sKL) divergences between the actual FDD and the Benford law distribution are given.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	***sKL***
$4.2 \leq MOS \leq 5$	0.309	0.183	0.121	0.096	0.093	0.059	0.052	0.046	0.041	0.004
$3.4 \leq MOS < 4.2$	0.313	0.180	0.121	0.099	0.089	0.059	0.050	0.046	0.043	0.004
$2.6 \leq MOS < 3.4$	0.316	0.180	0.121	0.099	0.090	0.058	0.049	0.045	0.043	0.004
$1.8 \leq MOS < 2.6$	0.322	0.177	0.118	0.098	0.096	0.056	0.046	0.043	0.043	0.008
$1 \leq MOS < 1.8$	0.331	0.173	0.114	0.098	0.102	0.054	0.043	0.042	0.044	0.014
Benford's law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0

Table 2. Mean FDD of Y directional gradient magnitudes in KoNViD-1k [22] with respect to different MOS intervals. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality. In the last column, the symmetric Kullback–Leibler (sKL) divergences between the actual FDD and the Benford law distribution are given.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	***sKL***
$4.2 \leq MOS \leq 5$	0.308	0.184	0.119	0.096	0.096	0.059	0.052	0.046	0.041	0.005
$3.4 \leq MOS < 4.2$	0.308	0.186	0.120	0.098	0.092	0.059	0.050	0.045	0.042	0.005
$2.6 \leq MOS < 3.4$	0.313	0.183	0.120	0.098	0.093	0.058	0.048	0.044	0.042	0.006
$1.8 \leq MOS < 2.6$	0.322	0.178	0.116	0.097	0.101	0.055	0.046	0.042	0.042	0.011
$1 \leq MOS < 1.8$	0.328	0.173	0.113	0.098	0.108	0.053	0.044	0.041	0.043	0.016
Benford's law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0

Table 3. Mean FDD of HLL wavelet coefficients in KoNViD-1k [22] with respect to different MOS intervals. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality. In the last column, the symmetric Kullback–Leibler (sKL) divergences between the actual FDD and the Benford law distribution are given.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	***sKL***
$4.2 \leq MOS \leq 5$	0.294	0.187	0.152	0.077	0.040	0.044	0.142	0.033	0.032	0.097
$3.4 \leq MOS < 4.2$	0.306	0.156	0.186	0.068	0.045	0.046	0.139	0.029	0.026	0.114
$2.6 \leq MOS < 3.4$	0.306	0.154	0.193	0.066	0.039	0.045	0.146	0.027	0.024	0.139
$1.8 \leq MOS < 2.6$	0.289	0.157	0.198	0.070	0.038	0.053	0.150	0.025	0.020	0.148
$1 \leq MOS < 1.8$	0.280	0.158	0.200	0.077	0.036	0.059	0.149	0.023	0.016	0.156
Benford's law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0

Table 4. Mean FDD of HHL wavelet coefficients in KoNViD-1k [22] with respect to different MOS intervals. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality. In the last column, the symmetric Kullback–Leibler (sKL) divergences between the actual FDD and the Benford law distribution are given.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	***sKL***
$4.2 \leq MOS \leq 5$	0.266	0.144	0.196	0.067	0.056	0.033	0.184	0.039	0.017	0.191
$3.4 \leq MOS < 4.2$	0.257	0.134	0.259	0.066	0.052	0.030	0.158	0.032	0.012	0.233
$2.6 \leq MOS < 3.4$	0.243	0.127	0.288	0.065	0.048	0.030	0.161	0.032	0.011	0.281
$1.8 \leq MOS < 2.6$	0.244	0.112	0.312	0.051	0.048	0.027	0.157	0.040	0.009	0.327
$1 \leq MOS < 1.8$	0.235	0.099	0.337	0.046	0.048	0.028	0.154	0.044	0.009	0.370
Benford's law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0

Table 5. Mean FDD of 3D DFT coefficients in KoNViD-1k [22] with respect to different MOS intervals. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality. In the last column, the symmetric Kullback–Leibler (sKL) divergences between the actual FDD and the Benford law distribution are given.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	***sKL***
$4.2 \leq MOS \leq 5$	0.306	0.170	0.120	0.095	0.079	0.069	0.060	0.053	0.048	6.18×10^{-4}
$3.4 \leq MOS < 4.2$	0.302	0.173	0.123	0.097	0.080	0.068	0.059	0.052	0.047	9.88×10^{-5}
$2.6 \leq MOS < 3.4$	0.294	0.172	0.125	0.100	0.082	0.069	0.060	0.052	0.046	4.45×10^{-4}
$1.8 \leq MOS < 2.6$	0.288	0.172	0.128	0.102	0.084	0.070	0.060	0.052	0.045	0.001
$1 \leq MOS < 1.8$	0.287	0.177	0.131	0.102	0.083	0.068	0.058	0.050	0.044	0.0011
Benford's law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0

Table 6. Mean FDD of 3D DCT coefficients in KoNViD-1k [22] with respect to different MOS intervals. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality. In the last column, the symmetric Kullback–Leibler (sKL) divergences between the actual FDD and the Benford law distribution are given.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	***sKL***
$4.2 \leq MOS \leq 5$	0.305	0.174	0.123	0.096	0.079	0.067	0.059	0.052	0.047	1.41×10^{-4}
$3.4 \leq MOS < 4.2$	0.302	0.175	0.124	0.097	0.079	0.067	0.059	0.052	0.046	2.92×10^{-5}
$2.6 \leq MOS < 3.4$	0.298	0.174	0.125	0.098	0.081	0.068	0.059	0.052	0.046	1.03×10^{-4}
$1.8 \leq MOS < 2.6$	0.295	0.174	0.126	0.099	0.081	0.069	0.059	0.052	0.046	2.42×10^{-4}
$1 \leq MOS < 1.8$	0.294	0.176	0.128	0.100	0.081	0.068	0.058	0.051	0.045	2.77×10^{-4}
Benford's law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0

Table 7. Mean FDD of higher-order singular values in KoNViD-1k [22] with respect to different MOS intervals. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality. In the last column, the symmetric Kullback–Leibler (sKL) divergences between the actual FDD and the Benford law distribution are given.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	***sKL***
$4.2 \leq MOS \leq 5$	0.303	0.175	0.124	0.096	0.079	0.067	0.058	0.052	0.046	2.75×10^{-5}
$3.4 \leq MOS < 4.2$	0.300	0.174	0.125	0.097	0.080	0.068	0.059	0.052	0.046	3.76×10^{-5}
$2.6 \leq MOS < 3.4$	0.297	0.174	0.125	0.098	0.081	0.068	0.059	0.052	0.046	1.02×10^{-4}
$1.8 \leq MOS < 2.6$	0.295	0.175	0.125	0.098	0.081	0.068	0.058	0.051	0.045	2.07×10^{-4}
$1 \leq MOS < 1.8$	0.295	0.177	0.128	0.099	0.081	0.068	0.058	0.051	0.045	2.05×10^{-4}
Benford's law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0

3.2. Perceptual Features

In our model, several perceptual features were also incorporated, which are consistent with human quality judgments [48]. Moreover, a subset of the presented perceptual features was also applied for no-reference image quality assessment in our previous work [16]:

1. Blur: This is the shape and area in an image that cannot be seen clearly because no distinct outline is present or an object is moving fast. Artifacts generated by blur usually result in the loss of details. Hereby, the amount of blur in an image heavily influences humans' quality perception. Due its low computational costs, we adopted the approach of Crété-Roffet et al. [49] to quantify the amount of blur in an image, which is based on the comparison between variations of adjacent pixels after low-pass filtering;
2. Colorfulness: There are more studies that suggest colorfulness as an important factor for human visual quality perception [48,50,51]. In our study, Hasler and Suesstrunk's model [52] was applied to measure colorfulness. Let' us denote with R, G, and B the red, green, and blue channels of an RGB image, respectively. Two matrices are derived for the color channels: $rg = R - G$ and $yb = \frac{1}{2}(R + G) - B$. Next, colorfulness (CF) is defined as:
$$CF = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + \frac{3}{10}\sqrt{\mu_{rg}^2 + \mu_{yb}^2}, \tag{12}$$
where σ^2 and μ stand for the variance and mean of their respective matrices. A video sequence's colorfulness is considered as the average value of individual frames' colorfulness;
3. Contrast: Perceptual image quality is strongly influenced by contrast, since humans' ability to distinguish objects from each other in an image heavily depends on it [53]. In [16], Matkovic et al.'s [54] global contrast factor (GCF) model was applied to quantify image contrast. However, GCF's computational cost is large, which makes it not feasible to measure a video sequence's contrast. That is why we adopted here the root-mean-squared (RMS) contrast for measuring the contrast of a video frame. RMS contrast is defined as the standard deviation of the pixel intensities [55]:
$$C_{RMS} = \sqrt{\frac{1}{M \cdot N}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(I_{i,j} - \bar{I})^2}, \tag{13}$$
where $I_{i,j}$ stands for the ith, jth pixel intensity of a 2D grayscale image I with size $M \times N$. A video sequence's contrast is considered as the average value of the video frames' contrast;
4. Dark channel feature: He et al. [56] investigated the properties of fog-free natural images. It was found that dark pixels are those pixels whose intensity values are close

to zero at least in one color channel within an image patch [57]. Based on this, a dark channel is defined as:

$$I^{dark}(x) = \min_{y \in \Omega(x)} (\min_{c \in \{R,G,B\}} I^c(y)), \tag{14}$$

where $I^c(y)$ denotes the intensity of a color channel (R, G, or B) and $\Omega(x)$ is an image patch centered on x. Based on the above definition, the dark channel feature (DCF) of an image is given as:

$$DCF = \frac{1}{||S||} \sum_{i \in S} \frac{I^{dark}(i)}{\sum_{c \in \{R,G,B\}} I^c(i)}, \tag{15}$$

where S stands for the area of the input image. A video sequence's DCF is considered as the average value of the individual video frames' DCF;

5. Entropy: The entropy of a digital image is a feature that gives information about the average content in an image. The concept of the entropy of a signal in general is very old. Namely, it comes from Shannon's theory of communication [58]. The entropy of a 2D grayscale image is given as:

$$E_I = -\sum_n p(n) \cdot \log_2 p(n), \tag{16}$$

where $p(n)$ stands for the empirical distribution of grayscale values in image I. The entropy of a video sequence is defined as the average of the video frames' entropy;

6. Mean of phase congruency: Phase congruency (PC) characterizes a digital image in the frequency domain. Phase congruency is given by the following equation:

$$PC_1(x) = \frac{|E(x)|}{\sum_n A_n(x)}, \tag{17}$$

where $E(x)$ corresponds to the energy of signal x and can be given as:

$$E(x) = |X(j\omega)|^2 \tag{18}$$

where $X(j\omega)$ is the Fourier transform of signal x and $A_n(x)$ denotes the nth Fourier amplitude of signal x. To incorporate noise compensation, Kovesi [59] modified the above definition of PC by adding weights for the frequency spread:

$$PC_2(x) = \frac{\sum_n W(x) \lfloor A_n(x) \Delta\phi_n(x) - T \rfloor}{\sum_n A_n(x) + \epsilon}, \tag{19}$$

where $W(x)$ is the weight function of the frequency spread, $\lfloor \cdot \rfloor$ stands for the floor function, T is an estimation of the noise level, and ϵ is a small constant to avoid division by zero. Moreover, $\phi_n(x)$ denotes the nth Fourier component at x and can be expressed as:

$$\Delta\phi_n(x) = cos(\phi_n(x) - \overline{\phi_n(x)}) - |sin(\phi_n(x) - \overline{\phi_n(x)})|, \tag{20}$$

where $\overline{\phi_n(x)}$ is the average phase at x. For a video sequence, the video frames' mean PC values are averaged to obtain a perceptual feature;

7. Spatial information: The gradient magnitude maps of each video frame were determined with the help of a Sobel filter, and the standard deviations of each Sobel map were taken. The spatial information (SI) of a video sequence is the average of the Sobel maps' standard deviations;
8. Temporal information: This characterizes the amount of temporal changes in a given video sequence [21]. In this study, the temporal information (TI) of a video sequence was considered as the mean of the pixelwise frame differences' standard deviations;

9. Natural image quality evaluator (NIQE): The NIQE [60] measures the distance between the natural scene statistics-based features extracted from an image and certain ideal features. In the case of the NIQE, the features are modeled as multidimensional Gaussian distributions. Specifically, the value given by the NIQE can be considered as the degree of deviation from naturalness of a digital image. In this study, the naturalness of a video sequence is characterized by the average of the video frames' NIQE values.

The average values of the above-described perceptual features with respect to five equal MOS intervals of KoNViD-1k [22] are given in Table 8. It can be observed that they are roughly proportional to the quality classes. With a properly chosen regression module, they can be good predictors of perceptual video quality.

Table 8. Mean of the perceptual features in KoNViD-1k [22] with respect to different MOS intervals. In KoNViD-1k [22], the lowest possible video quality is represented by $MOS = 1.0$, while $MOS = 5.0$ stands for the highest possible video quality.

	Blur	CF	Contrast	DCF	Entropy	PC	SI	TI	NIQE
$4.2 \leq MOS \leq 5$	0.309	0.229	0.211	0.197	7.027	0.019	83.478	0.034	3.745
$3.4 \leq MOS < 4.2$	0.371	0.196	0.223	0.244	7.103	0.017	70.850	0.067	3.802
$2.6 \leq MOS < 3.4$	0.423	0.193	0.226	0.223	6.800	0.013	59.306	0.081	4.163
$1.8 \leq MOS < 2.6$	0.458	0.198	0.188	0.153	6.260	0.007	42.072	0.077	4.888
$1 \leq MOS < 1.8$	0.451	0.213	0.158	0.098	5.577	0.007	34.056	0.081	5.356

4. Experimental Results and Analysis

In this section, our experimental results and analysis are presented. First, the applied benchmark VQA database is described in Section 4.1. Second, the evaluation metrics are given in Section 4.2. Third, the evaluation environment and implementation details are specified in Section 4.3. Fourth, a parameter study related to the proposed method is presented in Section 4.4 to reason about the design choices. Finally, a comparison to other state-of-the-art methods is presented in Section 4.5.

4.1. Databases

In our study, the KoNViD-1k [22] and LIVE VQC [23] benchmark VQA databases were used to carry out experiments and compare the proposed method to the state-of-the-art.

The videos of KoNViD-1k [22] were collected from the YFCC100M [61] database with respect to six predefined attributes: blur amount, colorfulness, contrast, spatial information, temporal information, and the natural image quality evaluator [60]. MOS values for each video were collected through a crowdsourcing process [62]. In this crowdsourcing process, 642 crowd workers from 64 countries participated, and they produced at least 50 judgments per video. In KoNViD-1k [22], $MOS = 1.0$ represents the lowest possible perceptual video quality, while $MOS = 5.0$ stands for the highest possible quality. The main characteristics of KoNViD-1k [22] are summarized in Table 9. The empirical distribution of the MOS values in KoNViD-1k [22] is depicted in Figure 5.

Similar to KoNViD-1k [22], LIVE VQC [23] contains authentically distorted video sequences with their corresponding perceptual quality scores. Specifically, it consists of 585 unique videos captured by 101 different video devices (mainly by smartphones). Moreover, the average length of the videos is 10 s. Similar to KoNViD-1k [22], the subjective quality scores were obtained in a crowdsourcing experiment where 4776 unique observers produced more than 205,000 opinion scores. The main characteristics of LIVE VQC [23] are summarized in Table 9. The empirical distribution of the MOS values in LIVE VQC [23] is depicted in Figure 6.

Table 9. Overview of the KoNViD-1k [22] and LIVE VQC [23] publicly available VQA databases.

Attribute	KoNViD-1k [22]	LIVE VQC [23]
Year	2017	2018
No. of sequences	1200	585
No. of scenes	1200	585
No. of devices	N/A	101
Device types	DSLR	smartphone
Distortion type	authentic	authentic
Duration	∼8 s	∼10 s
Resolution	960 × 540	320 × 240–1920 × 1080
Frame rate	30	N/A
Format	MPEG-4	N/A
Rating per video	50	200
MOS range	1.0–5.0	0.0–100.0

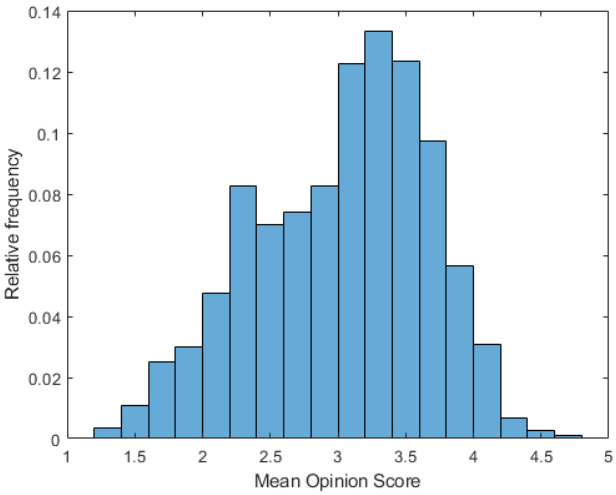


Figure 5. MOS distribution in KoNViD-1k [22].

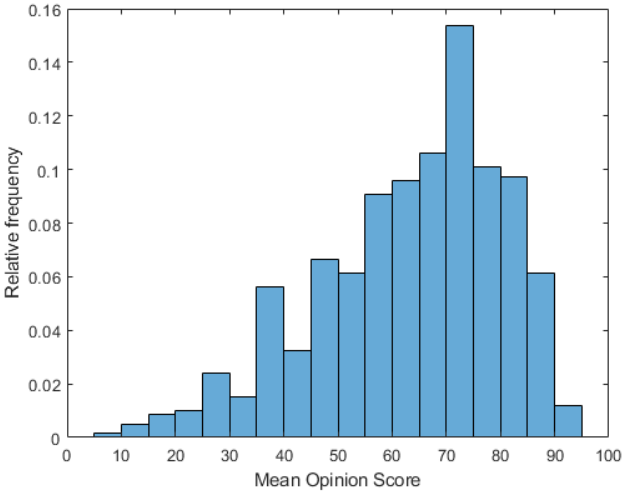


Figure 6. MOS distribution in LIVE VQC [23].

4.2. Evaluation Metrics

Similar to image quality assessment, the evaluation and performance ranking of NR-VQA algorithms rely on the measurement of the correlation between predicted and ground-truth perceptual quality scores. To this end, Pearson's linear correlation coefficient

(PLCC) and Spearman's rank-order correlation coefficient (SROCC) are widely applied in the literature [26]. The PLCC between the ground-truth and predicted quality scores can be defined as:

$$PLCC(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{21}$$

where N is the length of vectors $\mathbf{x}$ and $\mathbf{y}$, x_i and y_i are the ith elements of vectors $\mathbf{x}$ and $\mathbf{y}$, and finally, $\bar{x}$ and $\bar{y}$ stand for the mean of vectors $\mathbf{x}$ and $\mathbf{y}$. According to the recommendations of the Video Quality Expert Group [63], we adjusted the scaling and nonlinearity effect between the predicted and ground-truth scores by an $f(x)$ nonlinear transform, which is given by:

$$f(x) = \frac{\tau_1 - \tau_2}{1 + e^{-\frac{x - \tau_3}{|\tau_4|}}} + \tau_2, \tag{22}$$

where τ_1, τ_2, τ_3, and τ_4 are the fitting parameters. In contrast to the PLCC, the SROCC characterizes the monotonic relationship between the predicted and ground-truth quality scores and can be defined as:

$$SROCC(\mathbf{x}, \mathbf{y}) = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}, \tag{23}$$

where:

$$d_i = rank(x_i) - rank(y_i). \tag{24}$$

In this study, we report on the median values of the PLCC and SROCC after 1000 random training–testing splits.

4.3. Evaluation Environment and Implementation Details

To evaluate the proposed and the other state-of-the-art methods, KoNViD-1k [22] and LIVE VQC [23] were divided randomly into a training ($\sim$80%) and a test set ($\sim$20%). As already mentioned, the median PLCC and SROCC values are reported in this study, which were measured over 1000 random train–test splits. The computer configuration applied in our experiments is summarized in Table 10. Moreover, the proposed methods were implemented in MATLAB R2021a.

Table 10. Computer configuration applied in our experiments.

Computer model	STRIX Z270H Gaming
CPU	Intel(R) Core(TM) i7-7700K CPU 4.20 GHz (8 cores)
Memory	15 GB
GPU	Nvidia GeForce GTX 1080

4.4. Parameter Study

In this subsection, a parameter study is presented. Specifically, the performance of different FDD and perceptual features was examined with different regression modules, such as the SVR with linear and radial basis functions (RBFs), Gaussian process regression (GPR) with a rational quadratic kernel function, binary tree regression (BTR), and random forest regression (RFR). The results are summarized in Table 11. From these results, it can be clearly observed that the FDD features exhibited a rather weak or mediocre correlation with the ground-truth quality scores on KoNViD-1k [22], while considering all FDDs showed a rather strong correlation. In addition to this, the perceptual features also exhibited a strong correlation with the ground-truth quality scores on KoNViD-1k [22]. By fusing the FDDs and perceptual features together, a powerful feature vector can be obtained, which outperformed both the FDDs and perceptual features. Furthermore, GPR with a rational quadratic kernel function was the best-performing regression module, since it provided the highest correlation values almost in all cases. Based on these observations, we propose two NR-VQA methods, i.e., FDD-VQA and FDD + Perceptual-VQA, which

are compared to the state-of-the-art in the next subsection. FDD-VQA considers only the FDD-based feature vectors, while FDD + Perceptual-VQA fuses FDDs and perceptual features together. Both proposed methods rely on GPRs with rational quadratic kernel functions as the regression modules.

Table 11. Performance comparison of the FDD and perceptual feature vectors on KoNViD-1k [22]. Median SROCC values were measured over 1000 random train–test splits on KoNViD-1k [22].

Feature Vector	Linear SVR	RBF-SVR	GPR	BTR	RFR
FDD of X directional gradient magnitudes	0.402	0.419	0.432	0.223	0.218
FDD of Y directional gradient magnitudes	0.436	0.409	0.486	0.213	0.238
FDD of Z directional gradient magnitudes	0.394	0.359	0.386	0.206	0.183
FDD of HLL wavelet coefficients	0.320	0.302	0.347	0.152	0.171
FDD of LHL wavelet coefficients	0.279	0.382	0.412	0.201	0.202
FDD of HHL wavelet coefficients	0.425	0.493	0.503	0.323	0.328
FDD of LLH wavelet coefficients	0.338	0.387	0.414	0.220	0.237
FDD of HLH wavelet coefficients	0.347	0.394	0.421	0.237	0.250
FDD of LHH wavelet coefficients	0.316	0.412	0.428	0.229	0.246
FDD of HHH wavelet coefficients	0.449	0.479	0.498	0.323	0.304
FDD of 3D DFT coefficients	0.136	0.218	0.203	0.092	0.090
FDD of 3D DCT coefficients	0.135	0.190	0.207	0.132	0.092
FDD of higher-order singular values	0.156	0.117	0.144	0.097	0.091
Perceptual features	0.626	0.675	0.686	0.488	0.502
All FDDs	0.617	0.588	0.640	0.363	0.401
All FDDs + Perceptual	0.676	0.661	0.711	0.472	0.52

4.5. Comparison to the State-of-the-Art

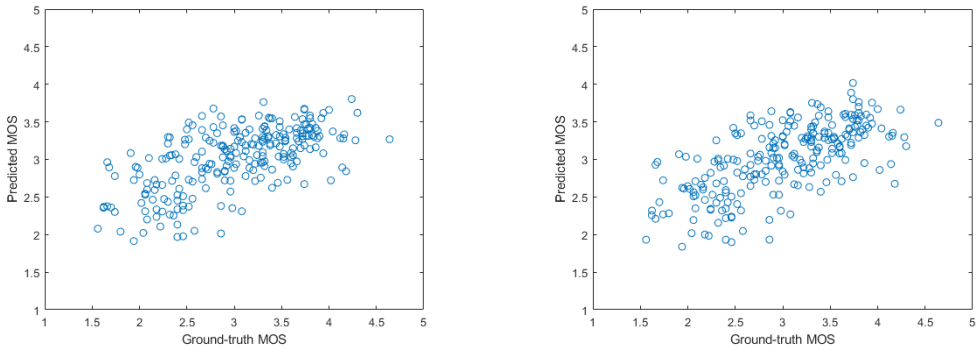
In this subsection, several NR-VQA methods, such as NVIE [64], V.BLIINDS [32], VIIDEO [65], 3D-MSCN [34], ST-Gabor [34], and 3D-MSCN + ST-Gabor [34], whose original source codes were made publicly available by the authors, are compared to the proposed *FDD-VQA* and *FDD + Perceptual-VQA* methods. These methods were evaluated exactly the same way as the proposed methods (described in Section 4.2). Moreover, the performance metrics of other state-of-the-art algorithms (FC model [66], STFC model [66], STS-SVR [67], STS-MLP [67], ChipQA [35]) were collected from the corresponding research studies.

The results for KoNViD-1k [22] are summarized in Table 12. It is clear from this table that the proposed NR-VQA approaches were able to provide competitive performance on a large, challenging VQA database both in terms of the PLCC and SROCC. Specifically, FDD-VQA, which solely relies on FDD feature vectors extracted from different domains (spatial, wavelet, Fourier, DCT, HOSVD), was able to outperform nine methods out of the examined eleven ones, while FDD + Perceptual-VQA outperformed all the considered state-of-the-art algorithms by a large margin. Figure 7 illustrates the scatter plots of the ground-truth MOS values against the predicted MOS values on the KoNViD-1k [22] test set.

The results for LIVE VQC [23] are summarized in Table 13. It can be seen that the proposed FDD-VQA and FDD + Perceptual-VQA methods were able to reach or outperform the state-of-the-art on LIVE VQC [23]. Specifically, FDD + Perceptual-VQA outperformed all the other examined NR-VQA algorithms, while FDD-VQA was able to reach the performance of the state-of-the-art.

Table 12. Comparison of *FDD-VQA* and *FDD + Perceptual-VQA* to the state-of-the-art on KoNViD-1k [22]. Median PLCC and SROCC values were measured over 1000 random train–test splits. The best results are in bold, while the second best results are underlined.

Method	PLCC	SROCC
NVIE [64]	0.404	0.333
V.BLIINDS [32]	0.661	0.694
VIIDEO [65]	0.301	0.299
3D-MSCN [34]	0.401	0.370
ST-Gabor [34]	0.639	0.628
3D-MSCN + ST-Gabor [34]	0.653	0.640
FC Model [66]	0.492	0.472
STFC Model [66]	0.639	0.606
STS-SVR [67]	0.680	0.673
STS-MLP [67]	0.407	0.420
ChipQA [35]	<u>0.697</u>	<u>0.694</u>
FDD-VQA	0.654	0.640
FDD + Perceptual-VQA	**0.716**	**0.711**



(**a**) (**b**)

Figure 7. Scatter plots of the ground-truth MOS against the predicted MOS of the proposed methods on the KoNViD-1k [22] test set. (**a**) FDD-VQA. (**b**) FDD + Perceptual-VQA.

Table 13. Comparison of *FDD-VQA* and *FDD + Perceptual-VQA* to the state-of-the-art on LIVE VQC [23]. Median PLCC and SROCC values were measured over 1000 random train–test splits. The best results are in bold, while the second best results are underlined. We denote by "-" when the data are not available.

Method	PLCC	SROCC
NVIE [64]	0.447	0.459
V.BLIINDS [32]	<u>0.690</u>	<u>0.703</u>
VIIDEO [65]	−0.006	−0.034
3D-MSCN [34]	0.502	0.510
ST-Gabor [34]	0.591	0.599
3D-MSCN + ST-Gabor [34]	0.675	0.677
FC Model [66]	-	-
STFC Model [66]	-	-
STS-SVR [67]	-	-
STS-MLP [67]	-	-
ChipQA [35]	0.669	0.697
FDD-VQA	0.623	0.630
FDD + Perceptual-VQA	**0.694**	**0.705**

To prove the significance of the presented results, one-sided *t*-tests were carried out on the 1000 SROCC values using a 95% confidence level. The results measured on KoNViD-1k are summarized in Table 14. From these results, it can be seen that the performance of FDD + Perceptual-VQA was statistically significantly better than those of the other examined state-of-the-art methods. Similarly, the results for LIVE VQC [23] are summed up in Table 15. It can be observed that the proposed FDD + Perceptual-VQA was statistically better than the other considered state-of-the-art methods except for V.BLIINDS [32], where no difference is exhibited.

Table 14. A one-sided *t*-test was carried among 1000 SROCC values measured on KoNViD-1k [22] using a 95% confidence level. In this table, "1" ("−1") denotes that the row algorithm is statistically better (worse) than the column algorithm.

	NVIE	**V.BLIINDS**	**VIIDEO**	**3D-MSCN**	**ST-Gabor**	**3D-MSCN + ST-Gabor**	**FDD + Perceptual-VQA**
NVIE	-	−1	1	−1	−1	−1	−1
V.BLIINDS	1	-	1	1	1	1	−1
VIIDEO	−1	−1	-	−1	−1	−1	−1
3D-MSCN	1	−1	1	-	−1	−1	−1
ST-Gabor	1	−1	1	1	-	−1	−1
3D-MSCN + ST-Gabor	1	−1	1	1	1	-	−1
FDD + Perceptual-VQA	1	1	1	1	1	1	-

Table 15. A one-sided *t*-test was carried among 1000 SROCC values measured on LIVE VQC [23] using a 95% confidence level. In this table, "1" ("−1") denotes that the row algorithm is statistically better (worse) than the column algorithm, and "0" stands for no statistical difference between the algorithms.

	NVIE	**V.BLIINDS**	**VIIDEO**	**3D-MSCN**	**ST-Gabor**	**3D-MSCN + ST-Gabor**	**FDD + Perceptual-VQA**
NVIE	-	−1	1	−1	−1	−1	−1
V.BLIINDS	1	-	1	1	1	1	0
VIIDEO	−1	−1	-	−1	−1	−1	−1
3D-MSCN	1	1	1	-	−1	−1	−1
ST-Gabor	1	−1	1	1	-	−1	−1
3D-MSCN + ST-Gabor	1	−1	1	1	1	-	−1
FDD + Perceptual-VQA	1	0	1	1	1	1	-

5. Conclusions

In this paper, we proposed a novel NR-VQA algorithm based on a set of novel quality-aware features, which relies on the FDDs of different domains (spatial, wavelet, DCT, DFT, HOSVD) and perceptual features. Specifically, we analyzed different FDD-based feature vectors in detail for NR-VQA. To this end, a detailed parameter study was established with respect to different domains and regression modules. It was demonstrated that state-of-the-art performance can be achieved in NR-VQA by considering only FDDs from different domains. Moreover, it was pointed out that fusing FDD and perceptual feature vectors together resulted in a powerful video representation for NR-VQA, which was able to outperform the state-of-the-art on two large authentically distorted VQA benchmark databases. Finally, the significance of the presented results was statistically proven with one-sided *t*-tests. Future work involves a real-time implementation of FDD feature extraction for NR-VQA on graphical processing units since many transformations and histogram calculations can be accelerated with them.

To facilitate the reproducibility of the presented results, the source code of the proposed method written in MATLAB R2021a is available at: https://github.com/Skythianos/Benford-VQA, accessed on 11 November 2021.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used were obtained from public open-source datasets from: 1. KoNViD-1k: http://database.mmsp-kn.de/konvid-1k-database.html (accessed on 11 November 2021); 2. LIVE VQC: https://live.ece.utexas.edu/research/LIVEVQC/index.html (accessed on 11 November 2021).

Acknowledgments: We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3D	three-dimensional
BTR	binary tree regression
C	contrast
CF	colorfulness
DCF	dark channel feature
DCT	discrete cosine transform
DFT	discrete Fourier transform
DSLR	digital single-lens reflex
DWT	discrete wavelet transform
FDD	first-digit distribution
FR	full-reference
FR-VQA	full-reference video quality assessment
GCF	global contrast factor
GPR	Gaussian process regression
HOSVD	higher-order singular-value decomposition
JPEG	Joint Photographic Experts Group
LIVE	Laboratory for Image and Video Engineering
MOS	mean opinion score
MPEG	Moving Picture Experts Group
MSCN	mean subtracted and contrast normalized
NIQE	natural image quality evaluator
NR	no-reference
NR-VQA	no-reference video quality assessment
PC	phase congruency
PLCC	Pearson's linear correlation coefficient
RBF	radial basis function
RFR	random forest regression
RMS	root mean square
RR	reduced-reference
RR-VQA	reduced-reference video quality assessment
SI	spatial information
SROCC	Spearman's rank-order correlation coefficient
SVR	support vector regressor
TI	temporal information
VQA	video quality assessment
VQC	video quality challenge
YFCC100M	Yahoo Flickr Creative Commons 100 Million

References

1. Index, C.V.N. Cisco visual networking index: Forecast and methodology 2015–2020. In *White Paper*; CISCO: San Jose, CA, USA, 2015.
2. Forecast, G. Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022. *Update* **2019**, *2017*, 2022.
3. Benford, F. The law of anomalous numbers. *Proc. Am. Philos. Soc.* **1938**, *78*, 551–572.
4. Fewster, R.M. A simple explanation of Benford's Law. *Am. Stat.* **2009**, *63*, 26–32. [CrossRef]
5. Özer, G.; Babacan, B. Benford's Law and Digital Analysis: Application on Turkish Banking Sector. *Bus. Econ. Res. J.* **2013**, *4*, 29–41. [CrossRef]

6. Hüllemann, S.; Schüpfer, G.; Mauch, J. Application of Benford's law: A valuable tool for detecting scientific papers with fabricated data? *Der Anaesthesist* **2017**, *66*, 795–802. [CrossRef] [PubMed]
7. Nye, J.; Moul, C. The political economy of numbers: On the application of Benford's law to international macroeconomic statistics. *BE J. Macroecon.* **2007**, *7*, 1–14. [CrossRef]
8. Gonzalez-Garcia, M.J.; Pastor, M.G.C. *Benford's Law and Macroeconomic Data Quality*; International Monetary Fund: Washington, DC, USA, 2009.
9. Rauch, B.; Göttsche, M.; Brähler, G.; Kronfeld, T. Deficit versus social statistics: Empirical evidence for the effectiveness of Benford's law. *Appl. Econ. Lett.* **2014**, *21*, 147–151. [CrossRef]
10. Jolion, J.M. Images and Benford's law. *J. Math. Imaging Vis.* **2001**, *14*, 73–81.
11. Pérez-González, F.; Heileman, G.L.; Abdallah, C.T. A generalization of Benford's law and its application to images. In Proceedings of the 2007 European Control Conference (ECC), Kos, Greece, 2–5 July 2007; pp. 3613–3619.
12. Pérez-González, F.; Heileman, G.L.; Abdallah, C.T. Benford's law in image processing. In Proceedings of the 2007 IEEE International Conference on Image Processing, San Antonio, TX, USA, 16–19 September 2007; Volume 1, pp. 405–408.
13. Fu, D.; Shi, Y.Q.; Su, W. A generalized Benford's law for JPEG coefficients and its applications in image forensics. In *Security, Steganography, and Watermarking of Multimedia Contents IX*; International Society for Optics and Photonics: Bellingham, WA, USA, 2007; Volume 6505, p. 65051L.
14. Andriotis, P.; Oikonomou, G.; Tryfonas, T. JPEG steganography detection with Benford's Law. *Digit. Investig.* **2013**, *9*, 246–257. [CrossRef]
15. Varga, D. Analysis of Benford's Law for No-Reference Quality Assessment of Natural, Screen-Content, and Synthetic Images. *Electronics* **2021**, *10*, 2378. [CrossRef]
16. Varga, D. No-reference image quality assessment based on the fusion of statistical and perceptual features. *J. Imaging* **2020**, *6*, 75. [CrossRef]
17. Hosu, V.; Lin, H.; Saupe, D. Expertise screening in crowdsourcing image quality. In Proceedings of the 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), Sardinia, Italy, 29–31 May 2018; pp. 1–6.
18. Hoßfeld, T.; Keimel, C.; Hirth, M.; Gardlo, B.; Habigt, J.; Diepold, K.; Tran-Gia, P. Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *IEEE Trans. Multimed.* **2013**, *16*, 541–558. [CrossRef]
19. ITU-R. *Methodology for the Subjective Assessment of the Quality of Television Pictures*; Recommendation ITU-R BT; International Telecommunication Union: Geneva, Switzerland, 2012; pp. 500–513.
20. International Telecommunication Union. *Methods for the Subjective Assessment of Video Quality Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment*; Series P: Terminals And Subjective and Objective Assessment Methods; International Telecommunication Union: Geneva, Switzerland, 2016.
21. ITU-T RECOMMENDATION. *Subjective Video Quality Assessment Methods for Multimedia Applications*; International Telecommunication Union: Geneva, Switzerland, 1999.
22. Hosu, V.; Hahn, F.; Jenadeleh, M.; Lin, H.; Men, H.; Szirányi, T.; Li, S.; Saupe, D. The Konstanz natural video database (KoNViD-1k). In Proceedings of the 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, 29 May–2 June 2017; pp. 1–6.
23. Sinno, Z.; Bovik, A.C. Large-scale study of perceptual video quality. *IEEE Trans. Image Process.* **2018**, *28*, 612–627. [CrossRef]
24. Winkler, S. Analysis of public image and video databases for quality assessment. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 616–625. [CrossRef]
25. Okarma, K. Image and video quality assessment with the use of various verification databases. In Proceedings of the New Electrical and Electronic Technologies and their Industrial Implementation, Zakopane, Poland, 18–21 June 2013; Volume 142.
26. Xu, L.; Lin, W.; Kuo, C.C.J. *Visual Quality Assessment by Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2015.
27. Winkler, S.; Mohandas, P. The evolution of video quality measurement: From PSNR to hybrid metrics. *IEEE Trans. Broadcast.* **2008**, *54*, 660–668. [CrossRef]
28. Argyropoulos, S.; Raake, A.; Garcia, M.N.; List, P. No-reference video quality assessment for SD and HD H. 264/AVC sequences based on continuous estimates of packet loss visibility. In Proceedings of the 2011 Third International Workshop on Quality of Multimedia Experience, Mechelen, Belgium, 7–9 September 2011; pp. 31–36.
29. Keimel, C.; Habigt, J.; Klimpke, M.; Diepold, K. Design of no-reference video quality metrics with multiway partial least squares regression. In Proceedings of the 2011 Third International Workshop on Quality of Multimedia Experience, Mechelen, Belgium, 7–9 September 2011; pp. 49–54.
30. Chen, Z.; Wu, D. Prediction of transmission distortion for wireless video communication: Analysis. *IEEE Trans. Image Process.* **2011**, *21*, 1123–1137. [CrossRef]
31. Pandremmenou, K.; Shahid, M.; Kondi, L.P.; Lövström, B. A no-reference bitstream-based perceptual model for video quality estimation of videos affected by coding artifacts and packet losses. In *Human Vision and Electronic Imaging XX*; International Society for Optics and Photonics: Bellingham, WA, USA, 2015; Volume 9394, p. 93941F.
32. Saad, M.A.; Bovik, A.C.; Charrier, C. Blind prediction of natural video quality. *IEEE Trans. Image Process.* **2014**, *23*, 1352–1365. [CrossRef] [PubMed]
33. Zhu, K.; Li, C.; Asari, V.; Saupe, D. No-reference video quality assessment based on artifact measurement and statistical analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 533–546. [CrossRef]

34. Dendi, S.V.R.; Channappayya, S.S. No-reference video quality assessment using natural spatiotemporal scene statistics. *IEEE Trans. Image Process.* **2020**, *29*, 5612–5624. [CrossRef]
35. Ebenezer, J.P.; Shang, Z.; Wu, Y.; Wei, H.; Bovik, A.C. No-reference video quality assessment using space-time chips. In Proceedings of the 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 21–24 September 2020; pp. 1–6.
36. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
37. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
38. Ren, Y.; Yu, X.; Chen, J.; Li, T.H.; Li, G. Deep image spatial transformation for person image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7690–7699.
39. Li, Y.; Po, L.M.; Cheung, C.H.; Xu, X.; Feng, L.; Yuan, F.; Cheung, K.W. No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *26*, 1044–1057. [CrossRef]
40. Ahn, S.; Lee, S. Deep blind video quality assessment based on temporal human perception. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 619–623.
41. Li, D.; Jiang, T.; Jiang, M. Unified quality assessment of in-the-wild videos with mixed datasets training. *Int. J. Comput. Vis.* **2021**, *129*, 1238–1257. [CrossRef]
42. Aqrawi, A.A.; Boe, T.H.; Barros, S. Detecting salt domes using a dip guided 3D Sobel seismic attribute. In *SEG Technical Program Expanded Abstracts 2011*; Society of Exploration Geophysicists: Tulsa, OK, USA, 2011; pp. 1014–1018.
43. Weeks, M.; Bayoumi, M. 3D discrete wavelet transform architectures. In Proceedings of the 1998 IEEE International Symposium on Circuits and Systems (Cat. No. 98CH36187), ISCAS'98, Monterey, CA, USA, 31 May–3 June 1998; Volume 4, pp. 57–60.
44. Heideman, M.T.; Johnson, D.H.; Burrus, C.S. Gauss and the history of the fast Fourier transform. *Arch. Hist. Exact Sci.* **1985**, *34*, 265–277. [CrossRef]
45. Baranyi, P.; Varlaki, P.; Szeidl, L.; Yam, Y. Definition of the HOSVD based canonical form of polytopic dynamic models. In Proceedings of the 2006 IEEE International Conference on Mechatronics, Budapest, Hungary, 3–5 July 2006; pp. 660–665.
46. Tucker, L.R. The extension of factor analysis to three-dimensional matrices. *Contrib. Math. Psychol.* **1964**, *110119*.
47. De Lathauwer, L.; De Moor, B.; Vandewalle, J. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **2000**, *21*, 1253–1278. [CrossRef]
48. Jenadeleh, M. Blind Image and Video Quality Assessment. Ph.D. Thesis, University of Konstanz, Konstanz, Germany, 2018.
49. Crete, F.; Dolmiere, T.; Ladret, P.; Nicolas, M. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *Human Vision and Electronic Imaging XII*; International Society for Optics and Photonics: Bellingham, WA, USA, 2007; Volume 6492, p. 64920I.
50. de Ridder, H. Naturalness and image quality: Saturation and lightness variation in color images of natural scenes. *J. Imaging Sci. Technol.* **1996**, *40*, 487–493.
51. Palus, H. Colorfulness of the image: Definition, computation, and properties. In *Lightmetry and Light and Optics in Biomedicine 2004*; International Society for Optics and Photonics: Bellingham, WA, USA, 2006; Volume 6158, p. 615805.
52. Hasler, D.; Suesstrunk, S.E. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*; International Society for Optics and Photonics: Bellingham, WA, USA, 2003; Volume 5007; pp. 87–95.
53. Segler, D.; Pettitt, G.; van Kessel, P. The importance of contrast and its effect on image quality. *SMPTE Motion Imaging J.* **2002**, *111*, 533–540. [CrossRef]
54. Matkovic, K.; Neumann, L.; Neumann, A.; Psik, T.; Purgathofer, W. *Global Contrast Factor-a New Approach to Image Contrast*; The Eurographics Association: Geneve, Switzerland, 2005; pp. 159–167.
55. Peli, E. Contrast in complex images. *JOSA A* **1990**, *7*, 2032–2040. [CrossRef]
56. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353.
57. Lee, S.; Yun, S.; Nam, J.H.; Won, C.S.; Jung, S.W. A review on dark channel prior based image dehazing algorithms. *EURASIP J. Image Video Process.* **2016**, *2016*, 1–23. [CrossRef]
58. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
59. Kovesi, P. Phase congruency detects corners and edges. In Proceedings of the Australian Pattern Recognition Society Conference, DICTA, Sydney, Australia, 10–12 December 2003; Volume 2003.
60. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* **2012**, *20*, 209–212. [CrossRef]
61. Thomee, B.; Shamma, D.A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; Li, L.J. YFCC100M: The new data in multimedia research. *Commun. ACM* **2016**, *59*, 64–73. [CrossRef]
62. Saupe, D.; Hahn, F.; Hosu, V.; Zingman, I.; Rana, M.; Li, S. Crowd workers proven useful: A comparative study of subjective video quality assessment. In Proceedings of the 8th International Conference on Quality of Multimedia Experience, QoMEX 2016, Lisbon, Portugal, 6–8 June 2016.

63. Rohaly, A.M.; Corriveau, P.J.; Libert, J.M.; Webster, A.A.; Baroncini, V.; Beerends, J.; Blin, J.L.; Contin, L.; Hamada, T.; Harrison, D.; et al. Video quality experts group: Current results and future directions. In *Visual Communications and Image Processing 2000*; International Society for Optics and Photonics: Bellingham, WA, USA, 2000; Volume 4067, pp. 742–753.
64. Mittal, A. Natural Scene Statistics-Based Blind Visual Quality Assessment in the Spatial Domain. Ph.D. Thesis, The University of Texas at Austin, Austin, TX, USA, 2013.
65. Mittal, A.; Saad, M.A.; Bovik, A.C. A completely blind video integrity oracle. *IEEE Trans. Image Process.* **2015**, *25*, 289–300. [CrossRef] [PubMed]
66. Men, H.; Lin, H.; Saupe, D. Spatiotemporal feature combination model for no-reference video quality assessment. In Proceedings of the 2018 Tenth international conference on quality of multimedia experience (QoMEX), Sardinia, Italy, 29–31 May 2018; pp. 1–3.
67. Yan, P.; Mou, X. No-reference video quality assessment based on perceptual features extracted from multi-directional video spatiotemporal slices images. In *Optoelectronic Imaging and Multimedia Technology V*; International Society for Optics and Photonics: Bellingham, WA, USA, 2018; Volume 10817, p. 108171D.