sensors

MDPI

Article

No-Reference Video Quality Assessment Using the Temporal Statistics of Global and Local Image Features

Domonkos Varga

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

Abstract: During acquisition, storage, and transmission, the quality of digital videos degrades significantly. Low-quality videos lead to the failure of many computer vision applications, such as object tracking or detection, intelligent surveillance, etc. Over the years, many different features have been developed to resolve the problem of no-reference video quality assessment (NR-VQA). In this paper, we propose a novel NR-VQA algorithm that integrates the fusion of temporal statistics of local and global image features with an ensemble learning framework in a single architecture. Namely, the temporal statistics of global features reflect all parts of the video frames, while the temporal statistics of local features reflect the details. Specifically, we apply a broad spectrum of statistics of local and global features to characterize the variety of possible video distortions. In order to study the effectiveness of the method introduced in this paper, we conducted experiments on two large benchmark databases, i.e., KoNViD-1k and LIVE VQC, which contain authentic distortions, and we compared it to 14 other well-known NR-VQA algorithms. The experimental results show that the proposed method is able to achieve greatly improved results on the considered benchmark datasets. Namely, the proposed method exhibits significant progress in performance over other recent NR-VQA approaches.

Keywords: no-reference video quality assessment; quality-aware features; multi-feature fusion

Citation: Varga, D. No-Reference Video Quality Assessment Using the Temporal Statistics of Global and Local Image Features. *Sensors* **2022**, *22*, 9696. https://doi.org/10.3390/s22249696

Academic Editor: Guangtao Zhai

Received: 30 October 2022
Accepted: 9 December 2022
Published: 10 December 2022

1. Introduction

The recent rise in video-driven data consumption has presented manufacturers and telecommunications service providers with the problem of providing improved video services [1]. Further, it has also created a compelling necessity to monitor and regulate video quality [2]. As a consequence, video quality assessment (VQA) has received more and more attention from both academia [3] and industry [4]. In numerous video processing activities, including video capture, compression, and transport, VQA—which seeks to anticipate the perceived quality of a video—is still a challenging task. Similarly to image quality assessment (IQA), VQA is also divided into two groups, i.e., subjective and objective, in the literature [5]. Subjective VQA involves laboratory and crowdsourcing experiments [6] for collecting quality ratings from human observers by presenting them with various video sequences. Further, objective VQA deals with mathematical and computational models that are able to predict digital videos' perceptual quality consistently with human quality perception. Although subjective VQA is more reliable than objective VQA, since it collects quality ratings directly from humans, at the same time, it is expensive and time-consuming [7]. This is why it cannot be applied in real-time systems, and objective VQA is a hot research topic. Traditionally, objective VQA methods are further divided in the literature depending on the availability of the reference pristine (distortion-free) videos [8]. Specifically, no-reference (NR) VQA methods have no access to the reference methods, while full-reference (FR) VQA methods have complete access to them. Reduced-reference (RR) VQA methods have partial information about the reference videos. In practice, NR-VQA is highly demanded, since reference videos are unavailable in many real-world applications [9].

Researchers of visual physiology have demonstrated that the human visual system (HVS) tends to produce an unconscious global impression about a scene [10]. Next, the HVS focuses on the local details step by step [11–13]. The main contributions of this study are as follows. Based on the previous point, we extract the temporal statistics of both local and global image features for NR-VQA. Namely, the temporal statistics of global features reflect all parts of the video frames, while the temporal statistics of local features reflect the details. Inspired by our previous work [14], we adapt the statistics of local feature descriptors extracted from filtered images for NR-VQA to compile video-level local feature vectors. Namely, several HVS-inspired filters, i.e., Bilaplacian, high-boost, and derivative filters, were introduced to enhance the statistical regularities of an image that influence human quality perception. Specifically, these HVS-inspired filters were first applied over the color channels of a video frame. Next, the statistics of FAST (features from accelerated segment test) [15] feature descriptors were used to compile frame-level features. Video-level features were obtained through the temporal pooling of frame-level features. Further, we propose an ensemble learning framework to integrate the predicted quality scores of several machine learning techniques for efficient quality estimation. Due to the previously mentioned innovations, our experimental results demonstrate that the performance of the proposed method surpasses that of other recently published NR-VQA methods on two large VQA benchmark databases, i.e., KoNViD-1k [16] and LIVE VQC [17], which contain authentically distorted video sequences.

The following is the paper's flow. Section 2 reviews related and previous work. The proposed method is discussed in Section 3. Subsequently, Section 4 describes our experimental results and a comparison with the state of the art. Our conclusion is in Section 5.

2. Literature Review

Recent NR-VQA techniques can be classified into two broad categories: (i) those that only take into account spatial image-level characteristics and (ii) those that also take into account the temporal information between a video's frames [18]. Further, the majority of many modern NR-VQA methods apply some kind of machine or deep learning technique.

Image-based NR-VQA techniques borrow many ideas from NR-IQA and analyze the natural scene statistics (NSS) for quality prediction. The assumption behind NSS is that natural scenes follow certain statistical regularities that are distorted in the presence of image noise [19]. In the case of video data, many NSS-based algorithms independently measure frame-by-frame deviations from the "natural" statistics [20–22]. In [23], five simple perceptual features (blurriness, contrast, colorfulness, spatial information, temporal information) were determined frame by frame and temporally pooled to construct a video-level feature vector, which was mapped onto perceptual quality scores with a trained support vector regressor (SVR) [24]. Other approaches also took temporal information into consideration in addition to temporal pooling [25]. For instance, the image-based metric was developed further by V-BLIINDS [26], which incorporated time–frequency and temporal motion information as well. In contrast, Yan et al. [27] extracted features, i.e., moments of feature maps, gradient magnitudes' joint distributions, filtering responses of Laplacians of a Gaussian, and motion energy, from multi-directional spatiotemporal slices and mapped them onto quality scores with either a shallow neural network or an SVR. Similarly, Lemesle et al. [28] combined frame-level and video-level features for NR-VQA. After testing a wide combination of features, the authors concluded that the histogram of oriented gradients [29], edge information, fast Fourier transform [30], blur, contrast, freeze, and temporal-information-based features were the most informative ones for predicting video quality without a reference. Instead of perceptual features, Wang and Li [31] devised a statistical model for the speed perception of the human visual system, which was utilized for the estimation of motion information and perceptual uncertainty. Contrarily, Hosu et al. [16] introduced several video-level perceptual features and mapped them onto perceptual quality scores with the help of an SVR [24].

Deep learning has recently been utilized for NR-VQA. One of the first methods utilizing deep learning was SACONVA [32], which extracted feature vectors from video data via a 3D shearlet transform [33]. Next, these features were mapped onto quality scores using logistic regression and a convolutional neural network (CNN). In contrast, Wang et al. [34] combined deep spatial and temporal features for perceptual quality prediction. Specifically, spatial features were obtained through the pooling of a CNN's activations. Further, the standard deviations of motion vectors were considered as temporal features. Next, two predictions were obtained from these two sets of features, and they were combined by using a Bayes classifier for video quality prediction. Agarla [35] proposed an approach in which the image quality attributes, i.e., sharpness, graininess, lightness, and color saturation, of video frames were estimated first by using the deep features of a CNN. Based on these attributes, frame-level quality scores were estimated. Finally, a recurrent neural network was trained for video quality estimation by using the previously predicted frame-level scores as training data. The two-level video quality model (TVLQM) proposed by Korhonen [36] first computed low-complexity features from the entire video sequence before the extraction of high-complexity features. Further, the author fused traditional hand-crafted temporal features with deep features extracted from a CNN, which was trained to predict digital images' perceptual quality. Similarly, Agarla et al. [37] extracted frame-level quality-aware features by using pretrained CNNs, but they introduced a temporal modeling block containing a recurrent neural network (RNN) [38] and a temporal hysteresis pooling for quality prediction. Chen et al. [39] also applied RNNs for NR-VQA. To be more specific, this method consisted of two steps: (i) learning of quality degradation and (ii) modeling of motion effects. Similarly to the previously mentioned algorithms, the authors used CNNs for deep feature extraction. Further, a hierarchical temporal model that included an RNN was introduced for temporal down-sampling and gathering of motion information. Li et al. [40] took a similar approach, but they used a gated recurrent unit (GRU) [41] that was trained on the deep features extracted from a ResNet [42] network for perceptual quality estimation. This method was further improved by Zhang and Wang [43] provided texture features aside from deep features. In contrast, Chen et al. [39] extracted motion information from different temporal frequencies and trained a hierarchical recurrent network for video quality estimation. Contrary to the previously mentioned approaches, Li et al. [44] experimented with the idea of a mixed-dataset training strategy to improve the performance of NR-VQA by increasing the size of the training database and to boost the generalization capability of the implemented model. Further, this model was trained by two different loss functions, i.e., monotonicity- and linearity-induced loss. In [45], the authors first implemented a visual attention module that obtained frame-level perceptual quality scores. Next, video quality predictions were obtained with the help of a structure imitating human visual and memory attention.

3. Proposed Method

The training and testing processes of the proposed method are summarized in Figures 1 and 2. In the training stage, the statistics of local and global image features were extracted from each frame of a video sequence found in the training database. Subsequently, these image statistics were temporally pooled together to compile a quality-aware feature vector that characterized a given video. Based on the extracted video-level feature vectors, several different machine learning models, i.e., a generalized additive model (GAM) [46], an LSBoost algorithm [47], a Gaussian process regressor (GPR) [48] with rational quadratic kernel function, a neural network (NN) with one hidden layer containing 10 neurons [49], an SVR with a radial basis function (RBF) [24], a binary decision tree (BDT) [50], and an extra tree (ET) [51], were trained for perceptual quality estimation. In the testing stage, these trained models were used to generate quality scores for a previously unseen video. The final quality score was obtained by taking the arithmetic mean of the models' scores. In Sections 3.1 and 3.2, the processes of the extraction of global and local features are given. Further, in an ablation study (Section 4.2), we provide proof that the pro-

posed ensemble framework results in improved performance compared to the performance of the individual regressors.

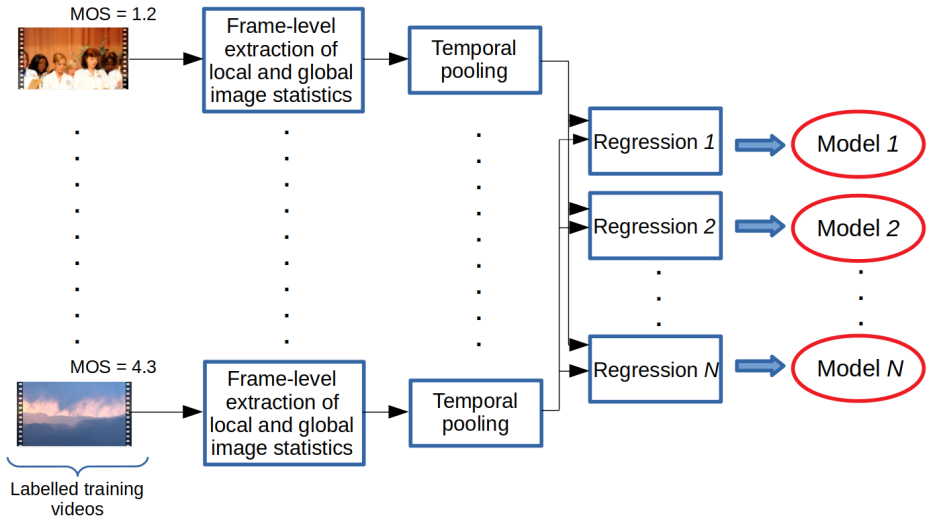


Figure 1. Training process of the proposed method. Video-level feature vectors are obtained from labeled training videos through the temporal pooling of local and global image statistics. Next, several regressors are trained.

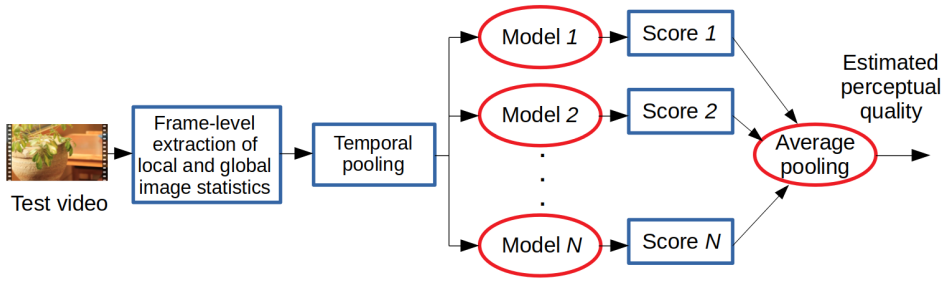


Figure 2. Testing of the proposed method. Video-level feature vectors are extracted from a test video through the temporal pooling of local and global image statistics. The scores of the trained regressors are fused together via average pooling to get an estimation of the perceptual quality.

3.1. Global Features

Many quality-aware features that characterize an image globally have been proposed in the literature in recent decades [52]. Due to their low computational complexities, BRISQUE [21], OG-IQA [53], SSEQ [54], and GM-LOG-BIQA [55] were utilized to compile video-level features through temporal pooling of their statistics. Specifically, BRISQUE [21] extracts features in the spatial domain. First of all, the mean subtracted normalized coefficient of an image is determined. Next, an asymmetric generalized Gaussian distribution (AGGD) is fitted to these coefficients. The parameters of the AGGD were considered quality-aware features. In contrast, OG-IQA [53] uses the variances in gradient magnitude, gradient orientation, and relative gradient magnitude maps as a feature vector. SSEQ [54] utilizes the spatial and spectral (discrete cosine transform coefficients) entropies of an image. GM-LOG-BIQA [55] compiles the joint distribution of the gradient magnitude and Laplacian features for quality-aware feature extraction. To define a global video-level feature vector, the previously mentioned quality-aware features were first determined for each video frame. Next, several well-known statistics, i.e., mean, median, standard deviation, entropy, skewness, and kurtosis, were extracted from a frame-level quality-aware feature. The arithmetic means of these statistics over time were considered as the video-

level quality-aware features. As a result, a vector with a length of 24 could be obtained for a single video sequence.

To boost the performance of the applied global features, the following set of perceptual features was also incorporated into our model.

1. Blur: This refers to the parts of an image that are out of focus. With too much blur, edges are no longer distinct. As a consequence, the amount of blur is an important element of human perceptual judgment. Due to its low computational complexity, the metric of Crété-Roffet et al. [56] was chosen in our model for the characterization of the amount of blur in a video frame. A video sequence's blur was defined as the average of all video frames' blur.
2. Colorfulness (CF): This is a characteristic of human visual perception that describes whether an image or image area seems to be more or less chromatic [57]. In [58], it was pointed out that humans tend to have a tendency toward more colorful scenes. In our model, we adopted the definition of colorfulness for a video frame proposed by Hasler and Suesstrunk [59]:

$$CF = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + \frac{3}{10}\sqrt{\mu_{rg}^2 + \mu_{yb}^2}, \tag{1}$$

where $rg = R - G$ and $yb = \frac{1}{2}(R + G) - B$. Further, R, G, and B denote the red, green, and blue color channels, respectively. The variables of μ and σ stand for the means and standard deviations of the matrices given in the subscripts, respectively. A video sequence's colorfulness was considered as the average of all video frames' colorfulness.
3. Vividness was suggested as a color attribute by Berns [60], and it describes the degree of departure of the color from a neutral black color. Berns' model can be expressed by the following formula:

$$V_B = \sqrt{(L^*)^2 + (a^*)^2 + (b^*)^2}, \tag{2}$$

where L^*, a^*, and b^* correspond to the color channels' values in the CIELAB color space [60,61]. In this study, the vividness of an image was defined by the average of all V_B values calculated from CIELAB's channels. As a quality-aware feature for a video sequence, the average of all video frames' vividness was taken.
4. The heaviness of a given color is also expressed with the help of the CIELAB space [62,63]:

$$H = 3.8 - 0.07 \cdot L^*. \tag{3}$$

In this study, the heaviness of an image was defined by the average of all H values calculated from CIELAB's channels. As a quality-aware feature for a video sequence, the average of all video frames' heaviness was taken.
5. Depth is also a color attribute, but it characterizes the degree of departure of a given color from a neutral white color, and in Berns' model [60], it is formally given as:

$$D_B = \sqrt{(100 - L^*)^2 + (a^*)^2 + (b^*)^2}. \tag{4}$$

In this study, the depth of an image was defined by the average of all D_B values calculated from CIELAB's channels. As a quality-aware feature for a video sequence, the average of all video frames' depth was taken.
6. The spatial information (SI) of a video frame is defined with the help of the non-maximum suppression (NMS) [64,65] algorithm. Namely, a video frame is characterized as the number of detected local extrema using three different T thresholds

($T = 1$, $T = 15$, and $T = 30$ were considered in this study). More specifically, the filtered video frame in which NMS is carried out is defined as follows:

$$F(x,y,T) = \begin{cases} 1, & \text{if } \forall_{(x,y)} I(x,y) > I(x+i,y+j) + T, \\ 1, & \text{else if } \forall_{(x,y)} I(x,y) < I(x+i,y+j) - T, \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $I(x,y)$ represents the value of pixel intensities at location (x,y). Further, $(i,j) \in \{(0,1),(0,-1),(1,0),(-1,0)\}$. In other words, the 3×3 neighborhood around (x,y) is considered. The SI of a video frame was defined as the entropy of the detected extremes' pixel intensities by using the three different previously mentioned thresholds. As a quality-aware feature for a video sequence, the average of all frames' SI was utilized.

7. Temporal information was defined by using the difference between two consecutive video frames. Namely, the standard deviations of all difference maps were determined, and their arithmetic mean was considered as a video-level quality-aware feature.
8. The color gradient magnitude (CGM) map of an RGB digital image is defined as

$$CGM(x) = \sum_{c \in (R,G,B)} \sqrt{(I_x^c(x))^2 + (I_y^c(x))^2}, \quad (6)$$

where the approximate directional derivatives of $I(x)$ in the horizontal and vertical directions are denoted by $I_x(x)$ and $I_y(x)$, respectively. A video frame was characterized by the mean of its CGM, while the average of all video frames' CGM means was considered as a quality-aware feature for a video sequence.
9. In addition to the mean of the CGM, the standard deviation of the CGM is also considered a quality-aware feature for a single video frame. As in the previous point, the average of all video frames' standard deviation was used to characterize the whole video sequence.
10. Sharpness determines the amount of detail in an image. It is most visible in image edges, and many approaches measure it with the step response. In our model, we estimated the sharpness of a video frame by using image gradients. Namely, the gradient magnitude map ($\mathbf{G}$) was calculated as

$$\mathbf{G} = \sqrt{(\mathbf{G}_x * \mathbf{I})^2 + (\mathbf{G}_y * \mathbf{I})^2}, \quad (7)$$

where $\mathbf{G}_x$ and $\mathbf{G}_y$ are horizontal and vertical Sobel operators, respectively. Further, $\mathbf{I}$ denotes an input grayscale image and $*$ stands for the convolution operator. The sharpness of image I is defined as the average value of the gradient magnitude map.
11. Michelson contrast: By definition, contrast corresponds to the difference in luminance that makes an object noticeable in an image [66]. Humans tend to appreciate images with higher contrast, since they can better distinguish between differences in intensity. In our model, we incorporated two different quantizations of contrast, i.e., Michelson and root mean square (RMS) contrast. The Michelson contrast of a still image is determined as follows:

$$C_{Michelson} = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}, \quad (8)$$

where I_{max} and I_{min} stand for the highest and lowest luminance, respectively. As a video perceptual feature, the average of all video frames' Michelson contrast was taken.
12. The RMS contrast of image with size $M \times N$ corresponds to the standard deviation of intensities [67]:

$$C_{RMS} = \sqrt{\frac{1}{M \cdot N} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} (I_{i,j} - \bar{I})^2}, \quad (9)$$

where $I_{i,j}$ denotes the intensity value at pixel position (i, j). Further, $\bar{I}$ stands for the arithmetic mean of all intensities. As a video perceptual feature, the average of all video frames' RMS contrast was taken.

13. The mean of an image gives the contribution sof individual pixel intensities for the entire image. Further, the mean is inversely proportional to the haze. In our study, the average of all video frames was considered as a quality-aware feature.
14. Entropy: This can be viewed as a measure of disorder in a digital image, and at the same time, it is a statistical feature that gives information about the average information content of an image [54]. Further, entropy tends to increase in an image as the intensity of noise or degradation levels increase [68]. An 8-bit-depth grayscale image's entropy (E) can be given as

$$E = -\sum_{n=0}^{255} p(n) \cdot \log_2(p(n)), \tag{10}$$

where $p(\cdot)$ corresponds to the image's normalized histogram count. In our model, a video sequence's entropy corresponds to the arithmetic mean of all video frames' entropy.

15. A perception-based image quality evaluator (PIQE) [69] is an opinion-unaware image quality estimator that does not require any training data. Further, it estimates perceptual quality only from salient image regions. First, an input image is divided into non-overlapping 16×16-sized blocks. The identification of salient blocks is carried out with the help of mean subtracted contrast normalized (MSCN) coefficients. Moreover, noise and artifact quantization are also carried out with MSCN coefficients. In our study, the average of all video frames' PIQE metrics was considered as a quality-aware feature.
16. The naturalness image quality evaluator (NIQE) [20] is also an opinion-unaware image quality estimator that needs no training data. Namely, it quantifies image quality as the distance between the NSS features of an input image and the NSS features of a model that was obtained from pristine (distortion-free) images. The applied NSS features are modeled as multidimensional Gaussian distributions. In our study, the average of all video frames' NIQE metrics was considered as a quality-aware feature.

3.2. Local Features

In our previous work, we empirically proved that the statistics of local feature descriptors are quality-aware features [14]. Further, if we apply certain human visual system (HVS)-inspired filters, dense feature vectors can be obtained. Influenced by our previous work, the following HVS-inspired image filters were applied: Bilaplacian filters, high-boost filters, and derivative filters. To be more specific, the Bilaplacian filters were motivated by the papers of Ghosh et al. [70,71], who demonstrated that the behavior of retinal ganglion cells' extended classical receptive field can be described by a combination of three zero-mean Gaussians at three different scales, which corresponds to the Bilaplacian of the Gaussian filter. Similarly to our previous work, the following Laplacian kernels are taken into consideration:

$$\mathbf{L}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \mathbf{L}_2 = \begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{pmatrix}, \mathbf{L}_3 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \tag{11}$$

$$\mathbf{L}_4 = \begin{pmatrix} -2 & 1 & -2 \\ 1 & 4 & 1 \\ -2 & 1 & -2 \end{pmatrix}, \mathbf{L}_5 = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}. \tag{12}$$

As the terminology indicates, a Bilaplacian kernel can be obtained through the convolution of two Laplacian kernels:

$$\mathbf{L}_{ij}^2 = \mathbf{L}_i * \mathbf{L}_j, \tag{13}$$

where the convolution operator is denoted by $*$. As in our previous study, $\mathbf{L}_{11}^2$, $\mathbf{L}_{22}^2$, $\mathbf{L}_{33}^2$, $\mathbf{L}_{44}^2$, $\mathbf{L}_{55}^2$, $\mathbf{L}_{13}^2$, and $\mathbf{L}_{24}^2$ Bilaplacian kernels were applied.

High-boost filtering was motivated by the property of the HVS that it is sensitive to the high-frequency regions of a natural scene [72]. In this paper, the following kernel was used:

$$\mathbf{H} = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{pmatrix}. \tag{14}$$

Since image distortions can occur at different scales, this filter was used 4 times in succession.

Derivative filters for visual quality assessment were used first by Li et al. [73], since statistical regularities of a natural scene could be extracted by them. In our study, the following convolution of two derivative kernels was applied:

$$\mathbf{D}_1 = \begin{pmatrix} -1 & 1 & -1 \\ 1 & 0 & 1 \\ -1 & 1 & -1 \end{pmatrix} * \begin{pmatrix} 1 & -1 & 1 \\ -1 & 0 & -1 \\ 1 & -1 & 1 \end{pmatrix}. \tag{15}$$

Since image distortions can occur at various scales of an image, $\mathbf{D}_2$, $\mathbf{D}_3$, $\mathbf{D}_4$, and $\mathbf{D}_5$ in sizes of 5×5, 7×7, 11×11, and 13×13 were also applied.

Using the previously described filters, the following set of kernels can be defined:

$$\mathcal{S} = \{\mathbf{L}_{11}^2, \mathbf{L}_{22}^2, \mathbf{L}_{33}^2, \mathbf{L}_{44}^2, \mathbf{L}_{55}^2, \mathbf{L}_{13}^2, \mathbf{L}_{24}^2, \mathbf{H}, \mathbf{H}^2, \mathbf{H}^3, \mathbf{H}^4, \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3, \mathbf{D}_4, \mathbf{D}_5\}. \tag{16}$$

All of the elements of the set defined by Equation (16) were applied to the Y, Cb, and Cr channels of an input RGB frame. The conversion from RGB to YCbCr color space could be performed by the following matrix equation [74]:

$$\begin{pmatrix} Y \\ Cb \\ Cr \end{pmatrix} = \begin{pmatrix} 0.2568 & 0.5041 & 0.0979 \\ -0.1482 & -0.2910 & 0.4392 \\ 0.4392 & -0.3678 & -0.0714 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{17}$$

As a result, $3 \times 7 = 21$ Bilaplacian, $3 \times 4 = 12$ high-boost, and $3 \times 5 = 15$ derivative feature maps could be obtained from an input video frame. Next, FAST keypoints [15] were detected on all feature maps. Further, all keypoints were described by their 5×5 neighborhood. Each keypoint was described by a feature vector that consisted of the mean, median, standard deviation, skewness, and kurtosis of the grayscale values found in the 5×5 neighborhood. The feature vectors that characterized a feature map were obtained by concatenating the keypoints' feature vectors. In our implementation, we set the number of keypoints to 50, since over this value, we did not experience any improvement in the performance on the KoNViD-1k [16] VQA benchmark database. As a result, a $3 \times 7 \times 50 \times 5 = 5250$ length feature vector from the Bilaplacian maps, $3 \times 4 \times 50 \times 5 = 3000$ length feature vector from the high-boost maps, and $3 \times 5 \times 50 \times 5 = 3750$ length feature vector from the derivative maps could be obtained. Similarly to the previously described global features, several statistics, i.e., mean, median, standard deviation, entropy, skewness, and kurtosis, were obtained from them to create a frame-level quality-aware feature. The arithmetic means of these statistics over time were considered as video-level quality-aware feature vectors. As a results, a vector of length 18 could be obtained for a single video sequence.

For an overview, we have provided a summary of the features introduced in our method in Table 1.

Table 1. Description of features introduced in our method.

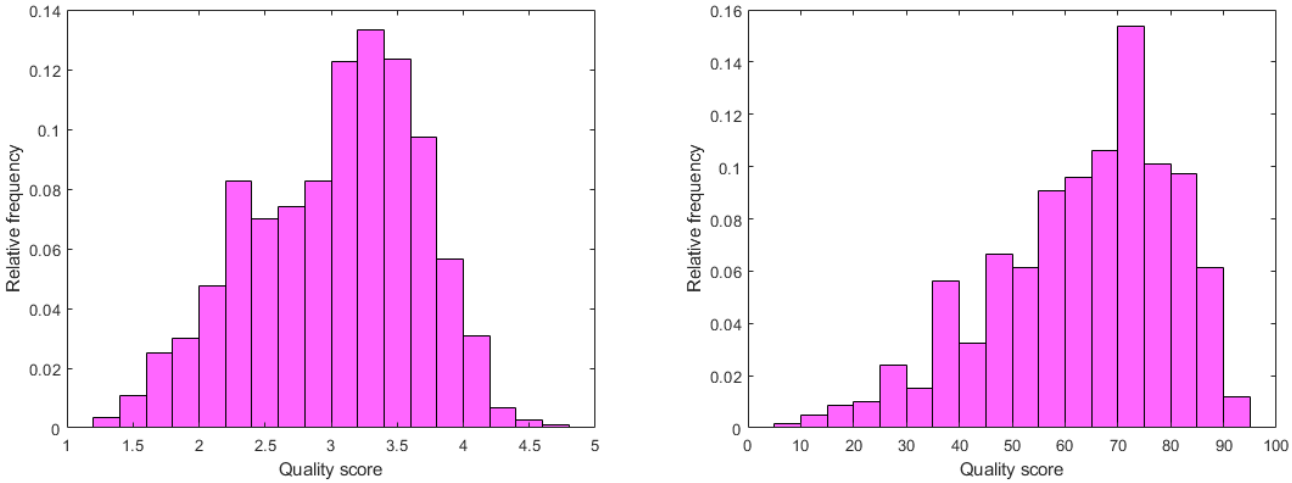
Feature Index	Description
f1–f6	Temporally pooled BRISQUE [21] statistics
f7–f12	Temporally pooled OG-IQA [53] statistics
f13–f18	Temporally pooled SSEQ [54] statistics
f19–f24	Temporally pooled GM-LOG-BIQA [55] statistics
f25–f40	Perceptual features
f41–f46	Temporally pooled Bilaplacian features' statistics
f47–f52	Temporally pooled high-boost features' statistics
f53–f58	Temporally pooled derivative features' statistics

4. Results

In this section, our experimental results are summarized. First, descriptions of the applied datasets and the evaluation protocol are given in Section 4.1. Next, a parameter study is used to justify the design choices of the proposed method in Section 4.2. Finally, the results of a comparison with the state-of-the-art methods are given in Section 4.3.

4.1. Datasets and Protocol

Experimental results and comparisons are presented on two large VQA databases that include digital videos with authentic distortions, i.e., KoNViD-1k [16] and LIVE VQC [17]. Hosu et al. [16] collected the 1200 videos found in KoNViD-1k [16] with an average length of 8 s from the YFCC100M database [75] with respect to several quality attributes, such as blur, colorfulness, contrast, spatial information, temporal information, and the numerical results of a natural image quality evaluator [20]. Quality scores for the selected videos were gathered in a crowdsourcing experiment involving 642 crowd workers from 64 countries. Further, quality scores were in the range of $[1.0, 5.0]$, where 1.0 denotes the lowest perceptual quality, while 5.0 is consistent with the highest perceptual quality. Unlike KoNViD-1k [16], LIVE VQC [17] includes 585 individual video sequences with an average length of 10 s, and the quality labels are in the range of $[0.0, 100.0]$. The evaluation of the videos was also carried out in a crowdsourcing process with 4776 unique observers. Figure 3 depicts the empirical distributions of quality scores in the databases that were used.



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

Figure 3. The empirical distributions of quality scores in the applied benchmark databases: (**a**) KoNViD-1k [16], (**b**) LIVE VQC [17]. The quality scores range from 1.0 to 5.0 in KoNViD-1k [16] and from 0.0 to 100.0 in LIVE VQC [17].

As recommended in the literature, a learning-based NR-VQA algorithm was trained on approximately 80% of the videos, and it was tested on the remaining 20% [76]. The performance of an NR-VQA method is characterized by the correlation strength between the predicted and ground-truth quality scores measured on the test set. To this end, Pearson's linear correlation coefficient (PLCC) and Spearman's rank order correlation coefficient (SROCC) are recommended. Following the guidelines of the Video Quality Expert Group [77], scaling and nonlinearity effects between predicted and ground-truth scores were adjusted by a nonlinear transform before the calculation of the PLCC. For the nonlinear regression of scores, the following function was adopted:

$$f(x) = \gamma_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\gamma_2 (x - \gamma_3))} \right) + \gamma_4 x + \gamma_5, \tag{18}$$

where γ_i $(i = 1, \cdots, 5)$ are the parameters to be fitted. The equations of the applied performance metrics are as follows:

$$PLCC = \frac{\sum_{i=1}^{N} (p_i - \bar{p})(m_i - \bar{m})}{\sqrt{\sum_{i=1}^{N} (p_i - \bar{p})^2 (m_i - \bar{m})^2}}, \tag{19}$$

where m_is are raw quality scores obtained from humans and p_is are the predictions provided by an NR-VQA algorithm. Further, $\bar{p}$ and $\bar{m}$ are mean values. The SROCC is defined as:

$$SROCC = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}, \tag{20}$$

where d_i refers to the difference between the ranks of both measures for observation i and N is the number of observations.

To ensure the stability of the numerical results, the medians of the PLCC and SROCC are reported in this study, and they were measured over 1000 random training–testing splits. Further, the proposed method was implemented in MATLAB R2022a, and the applied computer configuration is summarized in Table 2.

Table 2. Description of the computer configuration applied in our experiments.

Computer model	Z590 D
CPU	Intel(R) Core(TM) i7-11700F CPU 2.50 GHz (8 cores)
Memory	31.9 GB
GPU	Nvidia GeForce RTX 3090

4.2. Parameter Study

In this subsection, we justify the design choices of the proposed method. In Figure 4, a comparison of the performance of different regression techniques and strategies is depicted. The median PLCC and SROCC results were measured over 1000 random training–testing splits on KoNViD-1k [16]. From this figure, it can be seen that RBF SVR was the best single regressor, although the difference between RBF SVR and other single regressors was not too outstanding. More importantly, the mean or median pooling of the regressors' scores resulted in a significant performance improvement.

Figures 5 and 6 depict the PLCC and SROCC values of the different regression techniques and strategies in the form of box plots, respectively. On every box, the central mark represents the median value. Further, the bottom and top edges of the box correspond to the 25th and 75th percentiles, respectively. The whiskers continue to the most extreme data points that were not recognized as outliers, which are denoted by red '+' symbols. Figures 7 and 8 depict scatterplots of the ground truth versus the predicted scores on a KoNViD-1k [16] test set for each regression technique and strategy. Since the average pooling of the regressors' scores provided the best results according to our experiments on

KoNViD-1k [16], we applied this in our proposed method, which is referred to as FLG-VQA in the following, and in the comparison with other state-of-the-art methods.

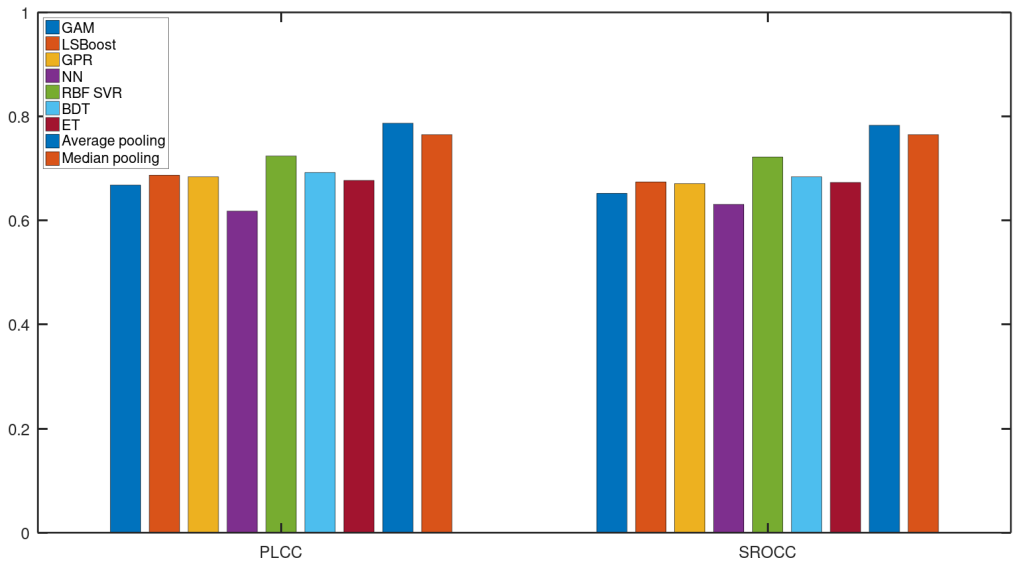


Figure 4. Performance comparison of different regression techniques (GAM, LSBoost, GPR, NN, RBF SVR, BDT, ET) and strategies (average pooling, median pooling) for the combination of individual regressors' results on KoNViD-1k [16]. The median PLCC and SROCC values, which were measured over 1000 random training–testing splits, are given.

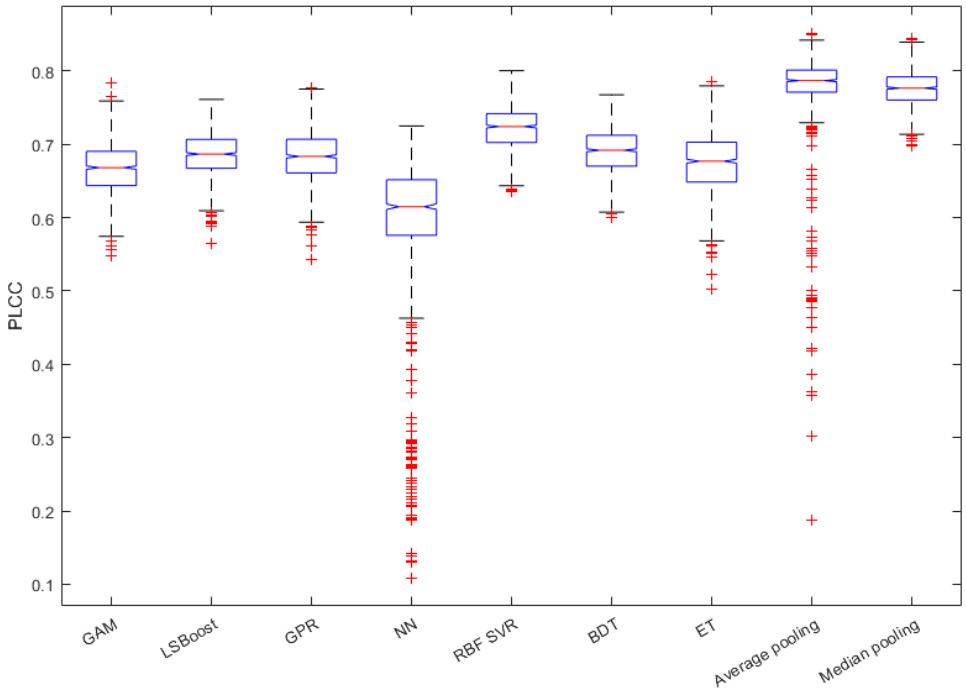


Figure 5. Box plots of PLCC values for different regression techniques and strategies. Measured over 1000 random training–testing splits on KoNViD-1k [16]. The bottom and top edges of each box correspond to the 25th and 75th percentiles, respectively. The whiskers continue to the most extreme data points that were not recognized as outliers, which are denoted by red '+' symbols.

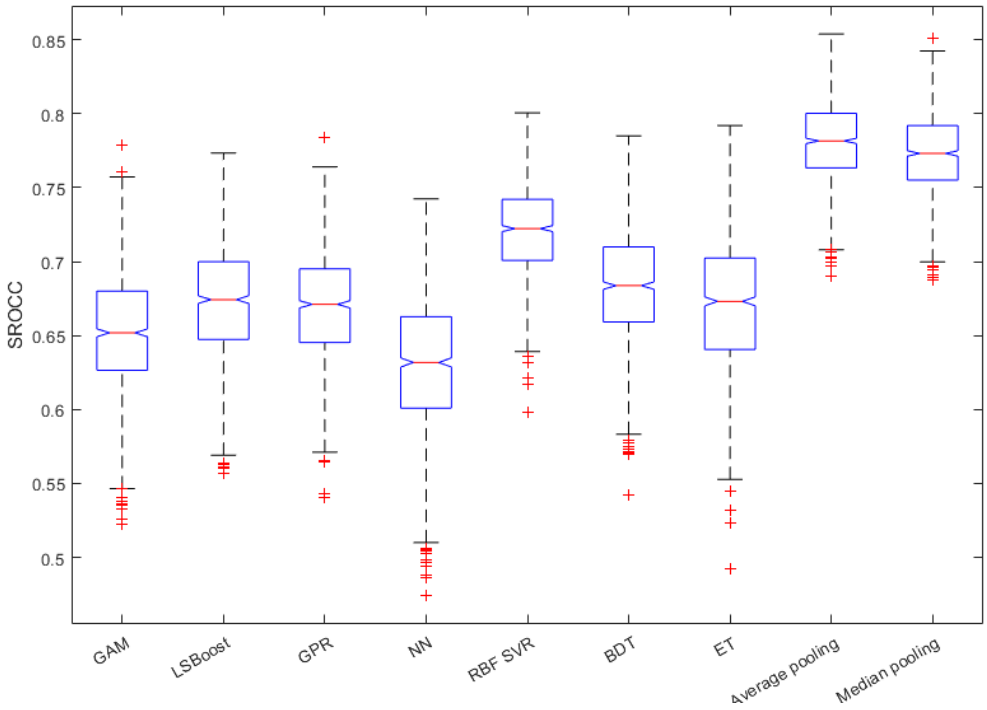


Figure 6. Box plots of SROCC values for different regression techniques and strategies. Measured over 1000 random training–testing splits on KoNViD-1k [16]. The bottom and top edges of each box correspond to the 25th and 75th percentiles, respectively. The whiskers continue to the most extreme data points that were not recognized as outliers, which are denoted by red '+' symbols.

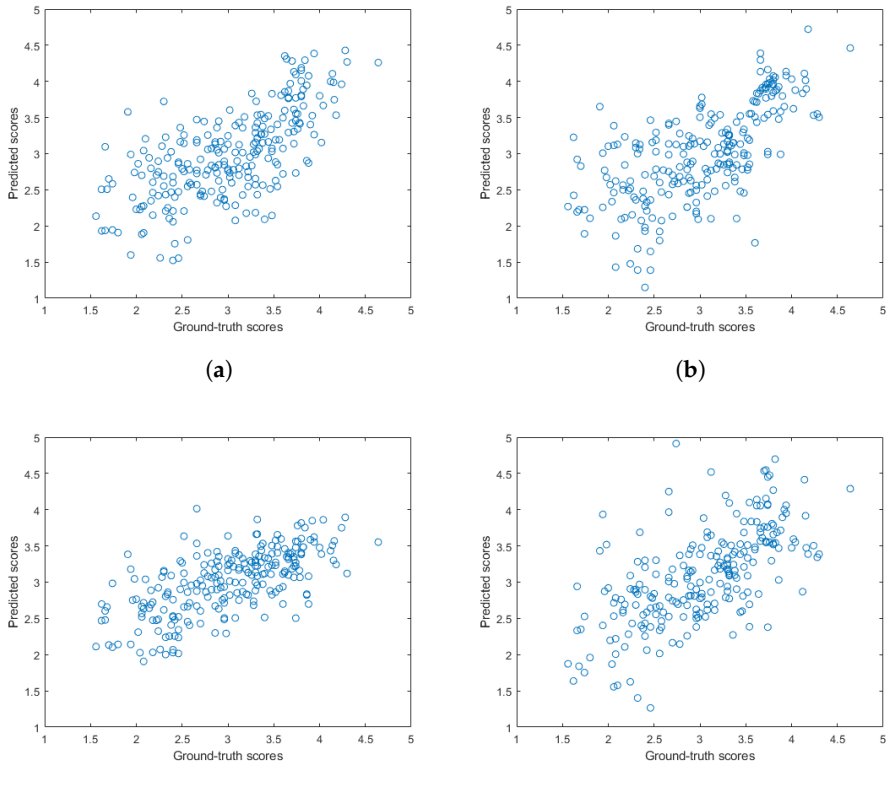


Figure 7. *Cont.*

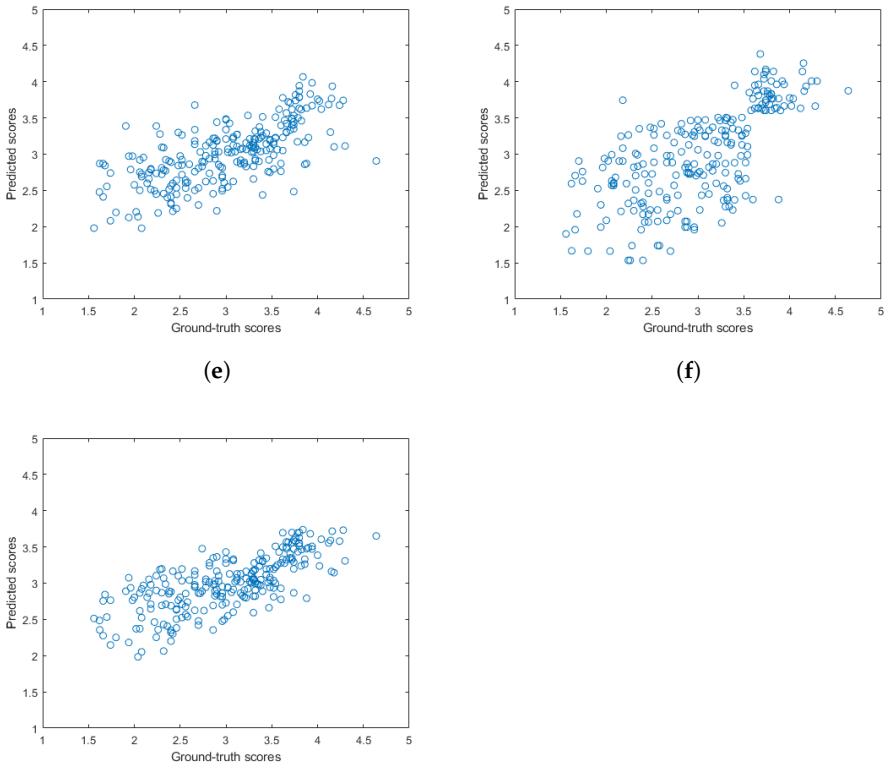


Figure 7. Scatterplots of the ground truth versus the predicted quality scores on a KoNViD-1k [16] test set for different regression techniques: (**a**) GAM, (**b**) LSBoost, (**c**) GPR, (**d**) NN, (**e**) RBF SVR, (**f**) BTR, and (**g**) ET.

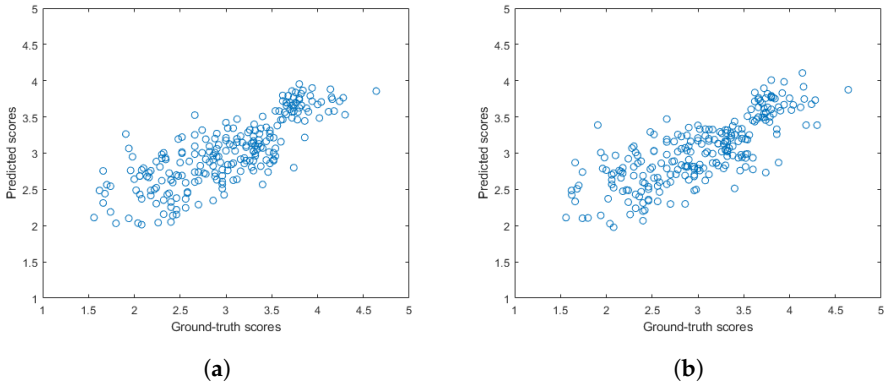


Figure 8. Scatterplots of the ground-truth versus the predicted quality scores on a KoNViD-1k [16] test set when using the pooling of individual regressors' scores as a regression strategy: (**a**) average pooling, (**b**) median pooling.

To demonstrate that all parts of the applied video-level feature vector in FLG-VQA are important and relevant, two additional experiments were also devised. First, the individual performance of each global and local feature was examined by using the evaluation protocol that was described in the previous subsection. The results of this experiment are summarized in Figure 9. As can be observed from these results, all global and local features were able to provide mediocre or rather strong results when considered on their own. It can be also observed that the temporal statistics of GM-LOG-BIQA [55] and the perceptual features provided the strongest individual performances, while the statistics of BRISQUE [21], SSEQ [54], and the high-boost filtered maps gave the weakest ones. The reason for this

is that BRISQUE and SSEQ [54] perform better on artificial image distortions, i.e., JPEG compression noise, than on authentic distortions [14], which are found in KoNViD-1k. Further, high-boost filtering is rather sensitive to high-frequency regions in a natural scene, which may restrict its performance on extremely different authentic distortions.

In the second experiment, we made an attempt to prove that all parts of the video-level feature vector are relevant. Namely, a given part of FLG-VQA's video-level feature vector with a length of 58 was eliminated, and then the performance of the remaining feature vector was examined. The results of the second experiment are summarized in Figure 10. From these results, it can be seen that the removal of any part of the feature vector resulted in a rather minor performance drop. Further, the removal of features that had strong individual performance did not result in a large decrease in the overall performance. Considering the experimental results in Figures 9 and 10 together, it seems to be justified that all parts of the proposed video-level feature vector are important and relevant. Further, it is worth considering global and local image statistics together in VQA.

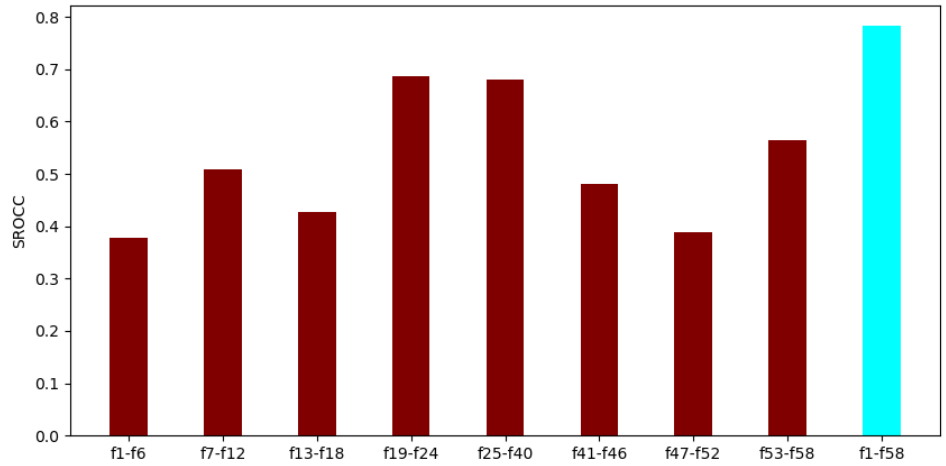


Figure 9. Performance comparison of the global and local features in FLG-VQA. The median SROCC values were measured on KoNViD-1k [16] over 1000 random training–testing splits. Table 1 gives information about the interpretation of the feature indices.

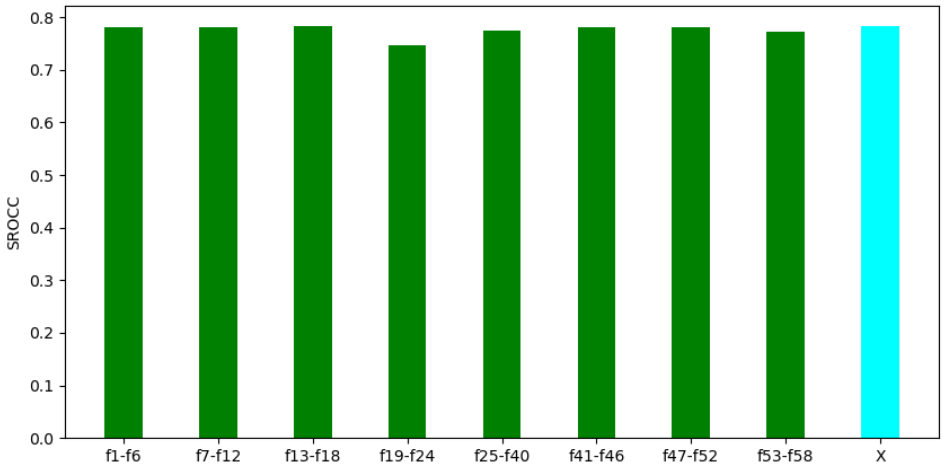


Figure 10. Performance of FLG-VQA in cases in which a part of the video-level feature vector was eliminated. The performance of the whole feature vector is denoted by 'X'. The median SROCC values were measured on KoNViD-1k [16] over 1000 random training–testing splits. Table 1 gives information about the interpretation of feature indices.

4.3. Comparison to the State-of-the-Art Methods

To verify the effectiveness of the proposed NR-VQA method, we compared the proposed algorithm with 10 other well-known methods, i.e., NVIE [78], V.BLIINDS [79], VIIDEO [80], 3D-MSCN [81], ST-Gabor [81], 3D-MSCN + ST-Gabor [81], FC Model [82], STFC Model [82], STS-SVR [27], STS-MLP [27], and ChipQA [83]. Specifically, the reported results of NVIE [78], V.BLIINDS [79], VIIDEO [80], 3D-MSCN [81], ST-Gabor [81], and 3D-MSCN + ST-Gabor [81] are based on our own experiments due to the availability of the original source codes of these methods. These methods were tested under exactly the same conditions as those of the proposed FLG-VQA. So, the median PLCC and SROCC values were measured after 1000 random training–testing splits, and approximately 80% of the videos were used for training, while the remaining ones were only applied in testing. The results of the other five NR-VQA methods were copied from their original publications. Further, Tu et al. [84] adapted two recently published deep-learning-based NR-IQA models, i.e., KonCept512 [85] and PaQ-2-PiQ [86], for NR-VQA. Their results, which were measured by the authors of [84], were also added to the presented comparison. Similarly to our evaluation protocol, the authors of [83,87] applied 1000 random training–testing splits and reported the median PLCC and SROCC values. Contrarily, Tu et al. [84] applied only 100 random splits, while the other papers used lower numbers of repetitions, i.e., 10 or 20. Moreover, the usual 80–20% split of the benchmark databases was used in all of the papers, since this choice is the most common and recommended for machine-learning-based methods in the literature.

The experimental results obtained on KoNViD-1k [16] and LIVE VQC [17] are summarized in Tables 3 and 4, respectively. Further, Table 5 summarizes the results of KoNViD-1k [16] and LIVE VQC [17] in the direct and weighted averages of the performance metrics. From the presented and summarized results, it can be observed that the proposed *FLG-VQA* was able to outperform the state-of-the-art methods by a large margin. For instance, the second best, ChipQA [87], was outperformed by approximately 0.02 in terms of both PLCC and SROCC on KoNViD-1k [16]. Similarly, on LIVE VQC [17], *FLG-VQA* provided results that were 0.01 and 0.02 higher than those of ChipQA [87] in terms of the PLCC and SROCC, respectively.

Table 3. Comparison of *FLG-VQA* with the state-of-the-art methods on KoNViD-1k [16]. The median PLCC and SROCC values were measured over 1000 random training–testing splits. The best results are in bold, while the second-best results are underlined.

Method	PLCC	SROCC
NVIE [78]	0.404	0.333
V.BLIINDS [79]	0.661	0.694
VIIDEO [80]	0.301	0.299
3D-MSCN [81]	0.401	0.370
ST-Gabor [81]	0.639	0.628
3D-MSCN + ST-Gabor [81]	0.653	0.640
FC Model [82]	0.492	0.472
STFC Model [82]	0.639	0.606
STS-SVR [27]	0.680	0.673
STS-MLP [27]	0.407	0.420
ChipQA-0 [83]	0.697	0.694
ChipQA [87]	<u>0.763</u>	<u>0.763</u>
KonCept512 [84,85]	0.749	0.735
PaQ-2-PiQ [84,86]	0.601	0.613
FLG-VQA	**0.787**	**0.783**

Table 4. Comparison of *FLG-VQA* with the state-of-the-art methods on LIVE VQC [17]. The median PLCC and SROCC values were measured over 1000 random training–testing splits. The best results are in bold, while the second-best results are underlined. We indicate with "-" when the data are not available.

Method	PLCC	SROCC
NVIE [78]	0.447	0.459
V.BLIINDS [79]	0.690	0.703
VIIDEO [80]	−0.006	−0.034
3D-MSCN [81]	0.502	0.510
ST-Gabor [81]	0.591	0.599
3D-MSCN + ST-Gabor [81]	0.675	0.677
FC Model [82]	-	-
STFC Model [82]	-	-
STS-SVR [27]	-	-
STS-MLP [27]	-	-
ChipQA-0 [83]	0.669	0.697
ChipQA [87]	0.723	<u>0.719</u>
KonCept512 [84,85]	<u>0.728</u>	0.665
PaQ-2-PiQ [84,86]	0.668	0.644
FLG-VQA	**0.733**	**0.731**

Table 5. Comparison of *FLG-VQA* with the state-of-the-art methods using the direct and weighted averages of the PLCC and SROCC values measured on the KoNViD-1k [16] and LIVE VQC databases [17].

	Direct Average		Weighted Average	
Method	**PLCC**	**SROCC**	**PLCC**	**SROCC**
NVIE [78]	0.426	0.396	0.418	0.374
V.BLIINDS [79]	0.676	0.698	0.671	0.697
VIIDEO [80]	0.148	0.133	0.200	0.190
3D-MSCN [81]	0.452	0.440	0.434	0.416
ST-Gabor [81]	0.615	0.613	0.623	0.618
3D-MSCN + ST-Gabor [81]	0.664	0.659	0.660	0.652
FC Model [82]	-	-	-	-
STFC Model [82]	-	-	-	-
STS-SVR [27]	-	-	-	-
STS-MLP [27]	-	-	-	-
ChipQA-0 [83]	0.683	0.696	0.688	0.695
ChipQA [87]	<u>0.743</u>	<u>0.741</u>	<u>0.750</u>	<u>0.749</u>
KonCept512 [84,85]	0.739	0.700	0.742	0.712
PaQ-2-PiQ [84,86]	0.635	0.629	0.623	0.623
FLG-VQA	**0.760**	**0.757**	**0.769**	**0.766**

5. Conclusions

NR-VQA, which has a high accuracy, has tremendous significance in many real-world applications. Specifically, a diverse set of local and global image features' statistics was proposed and applied with an ensemble learning framework to obtain a perceptual quality estimator. The main consideration behind this framework was that the HVS first produces an unconscious global impression of a visual scene. Next, the HVS turns its attention to fine local details. Many quality-aware features that characterize images globally have been proposed

over recent decades. We chose four of them to compile their statistics over time. Further, these statistics were boosted with several perceptual features. Moreover, local statistics were also derived with the help of three HVS-inspired filters (Bilaplacian, high-boost, and derivative filters) and the FAST keypoint detector to obtain dense frame-level feature vectors. The statistics of these dense vectors over time were considered as quality-aware features. After the fusion of the global and local statistics, an ensemble learning framework was used to map them onto perceptual quality scores. The proposed method was compared with 12 other recently published NR-VQA algorithms on the KoNViD-1k and LIVE VQC benchmark datasets. Our method's superiority in performance was demonstrated.

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: 1. KoNViD-1k: http://database.mmsp-kn.de/konvid-1k-database.html (accessed on 16 April 2022), 2. LIVE VQC: https://live.ece.utexas.edu/research/LIVEVQC/index.html (accessed on 16 April 2022).

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

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AGGD	asymmetric generalized Gaussian distribution
BDT	binary decision tree
CF	colorfulness
CGM	color gradient magnitude
CNN	convolutional neural network
E	entropy
ET	extra tree
FR-VQA	full-reference video quality assessment
GAM	generalized additive model
GPR	Gaussian process regressor
GRU	gated recurrent unit
HVS	human visual system
IQA	image quality assessment
LIVE	Laboratory for Image and Video Engineering
MOS	mean opinion score
NIQE	naturalness image quality evaluator
NMS	non-maximal suppression
NN	neural network
NR-IQA	no-reference image quality assessment
NR-VQA	no-reference video quality assessment
NSS	natural scene statistics
PIQE	perception-based image quality evaluator
PLCC	Pearson's linear correlation coefficient
RBF	radial basis function
RMS	root mean square
RNN	recurrent neural network
RR-VQA	reduced-reference video quality assessment
SI	spatial information
SROCC	Spearman's rank order correlation coefficient
SVR	support vector regressor
VQA	video quality assessment
VQC	video quality challenge

References

1. Hewage, C.T.; Ahmad, A.; Mallikarachchi, T.; Barman, N.; Martini, M.G. Measuring, modelling and Integrating Time-varying Video Quality in End-to-End Multimedia Service Delivery: A Review and Open Challenges. *IEEE Access* **2022**, *10*, 60267–60293. [CrossRef]
2. 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 QoMEX 2016: 8th International Conference on Quality of Multimedia Experience, Lisbon, Portugal, 6–8 June 2016.
3. Men, H.; Hosu, V.; Lin, H.; Bruhn, A.; Saupe, D. Subjective annotation for a frame interpolation benchmark using artefact amplification. *Qual. User Exp.* **2020**, *5*, 8. [CrossRef]
4. Brunnstrom, K.; Hands, D.; Speranza, F.; Webster, A. VQEG validation and ITU standardization of objective perceptual video quality metrics [standards in a nutshell]. *IEEE Signal Process. Mag.* **2009**, *26*, 96–101. [CrossRef]
5. Winkler, S. Video quality measurement standards—Current status and trends. In Proceedings of the 2009 IEEE 7th International Conference on Information, Communications and Signal Processing (ICICS), Macau, China, 8–10 December 2009; pp. 1–5.
6. Gadiraju, U.; Möller, S.; Nöllenburg, M.; Saupe, D.; Egger-Lampl, S.; Archambault, D.; Fisher, B. Crowdsourcing versus the laboratory: Towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 6–26.
7. Wit, M.T.; Wit, R.M.; Wit, N.B.; Ribback, R.; Iqu, K.R. 5G Experimentation Environment for 3rd Party Media Services D2. 9 Continuous QoS/QoE Monitoring Engine Development-Initial. 2022. Available online: https://www.5gmediahub.eu/wp-content/uploads/2022/06/D2.9_submitted.pdf (accessed on 29 October 2022).
8. Shahid, M.; Rossholm, A.; Lövström, B.; Zepernick, H.J. No-reference image and video quality assessment: A classification and review of recent approaches. *EURASIP J. Image Video Process.* **2014**, *2014*, 40. [CrossRef]
9. Ghadiyaram, D.; Chen, C.; Inguva, S.; Kokaram, A. A no-reference video quality predictor for compression and scaling artifacts. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3445–3449.
10. De Cesarei, A.; Loftus, G.R. Global and local vision in natural scene identification. *Psychon. Bull. Rev.* **2011**, *18*, 840–847. [CrossRef]
11. Bae, S.H.; Kim, M. A novel image quality assessment with globally and locally consilient visual quality perception. *IEEE Trans. Image Process.* **2016**, *25*, 2392–2406. [CrossRef]
12. Wang, H.; Qu, H.; Xu, J.; Wang, J.; Wei, Y.; Zhang, Z. Combining Statistical Features and Local Pattern Features for Texture Image Retrieval. *IEEE Access* **2020**, *8*, 222611–222624. [CrossRef]
13. Chang, H.w.; Du, C.Y.; Bi, X.D.; Chen, K.; Wang, M.H. Lg-Iqa: Integration of Local and Global Features for No-Reference Image Quality Assessment. *Displays* **2022**, *75*, 102334. *Available at SSRN 4108605* . [CrossRef]
14. Varga, D. A Human Visual System Inspired No-Reference Image Quality Assessment Method Based on Local Feature Descriptors. *Sensors* **2022**, *22*, 6775. [CrossRef] [PubMed]
15. Rosten, E.; Drummond, T. Fusing points and lines for high performance tracking. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Beijing, China, 17–21 October 2005; Volume 2, pp. 1508–1515.
16. 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 IEEE Ninth International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, 31 May–2 June 2017; pp. 1–6.
17. Sinno, Z.; Bovik, A.C. Large-scale study of perceptual video quality. *IEEE Trans. Image Process.* **2018**, *28*, 612–627. [CrossRef]
18. Kossi, K.; Coulombe, S.; Desrosiers, C.; Gagnon, G. No-reference video quality assessment using distortion learning and temporal attention. *IEEE Access* **2022**, *10*, 41010–41022. [CrossRef]
19. Srivastava, A.; Lee, A.B.; Simoncelli, E.P.; Zhu, S.C. On advances in statistical modeling of natural images. *J. Math. Imaging Vis.* **2003**, *18*, 17–33. [CrossRef]
20. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* **2012**, *20*, 209–212. [CrossRef]
21. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef] [PubMed]
22. Kundu, D.; Ghadiyaram, D.; Bovik, A.C.; Evans, B.L. No-reference quality assessment of tone-mapped HDR pictures. *IEEE Trans. Image Process.* **2017**, *26*, 2957–2971. [CrossRef]
23. Men, H.; Lin, H.; Saupe, D. Empirical evaluation of no-reference VQA methods on a natural video quality database. In Proceedings of the 2017 IEEE Ninth International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, 31 May–2 June 2017; pp. 1–3.
24. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
25. Xu, J.; Ye, P.; Liu, Y.; Doermann, D. No-reference video quality assessment via feature learning. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 491–495.
26. Saad, M.A.; Bovik, A.C. Blind quality assessment of videos using a model of natural scene statistics and motion coherency. In Proceedings of the IEEE 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove, CA, USA, 4–7 November 2012; pp. 332–336.

27. Yan, P.; Mou, X. No-reference video quality assessment based on perceptual features extracted from multi-directional video spatiotemporal slices images. In Proceedings of the Optoelectronic Imaging and Multimedia Technology V, International Society for Optics and Photonics, Beijing, China, 11–13 October 2018; Volume 10817, pp. 335–344.
28. Lemesle, A.; Marion, A.; Roux, L.; Gouaillard, A. NARVAL: A no-reference video quality tool for real-time communications. *Electron. Imaging* **2019**, *2019*, 213-1–213-7. [CrossRef]
29. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
30. Nussbaumer, H.J. The fast Fourier transform. In *Fast Fourier Transform and Convolution Algorithms*; Springer: Berlin/Heidelberg, Germany, 1981; pp. 80–111.
31. Wang, Z.; Li, Q. Video quality assessment using a statistical model of human visual speed perception. *JOSA A* **2007**, *24*, B61–B69. [CrossRef]
32. 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]
33. Lim, W.Q. The discrete shearlet transform: A new directional transform and compactly supported shearlet frames. *IEEE Trans. Image Process.* **2010**, *19*, 1166–1180.
34. Wang, C.; Su, L.; Zhang, W. COME for no-reference video quality assessment. In Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, USA, 10–12 April 2018; pp. 232–237.
35. Agarla, M.; Celona, L.; Schettini, R. No-reference quality assessment of in-capture distorted videos. *J. Imaging* **2020**, *6*, 74. [CrossRef] [PubMed]
36. Korhonen, J. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.* **2019**, *28*, 5923–5938. [CrossRef] [PubMed]
37. Agarla, M.; Celona, L.; Schettini, R. An Efficient Method for No-Reference Video Quality Assessment. *J. Imaging* **2021**, *7*, 55. [CrossRef] [PubMed]
38. Dupond, S. A thorough review on the current advance of neural network structures. *Annu. Rev. Control.* **2019**, *14*, 200–230.
39. Chen, P.; Li, L.; Ma, L.; Wu, J.; Shi, G. RIRNet: Recurrent-in-recurrent network for video quality assessment. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 834–842.
40. Li, D.; Jiang, T.; Jiang, M. Quality assessment of in-the-wild videos. In Proceedings of the 27th ACM International Conference on Multimedia, Nica, France, 21–25 October 2019; pp. 2351–2359.
41. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Amsterdam, The Netherlands, 8–10 October 2016; pp. 770–778.
43. Zhang, A.X.; Wang, Y.G. Texture Information Boosts Video Quality Assessment. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23–27 May 2022; pp. 2050–2054.
44. 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]
45. Guan, X.; Li, F.; Zhang, Y.; Cosman, P.C. End-to-End Blind Video Quality Assessment Based on Visual and Memory Attention Modeling. *IEEE Trans. Multimed.* **2022**, 1–16. [CrossRef]
46. Lou, Y.; Caruana, R.; Gehrke, J. Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 150–158.
47. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Seeger, M. Gaussian processes for machine learning. *Int. J. Neural Syst.* **2004**, *14*, 69–106. [CrossRef]
49. Wright, S.; Nocedal, J. Numerical optimization. *Springer Sci.* **1999**, *35*, 7.
50. Loh, W.Y. Regression tress with unbiased variable selection and interaction detection. *Stat. Sin.* **2002**, *12*, 361–386.
51. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
52. Zhu, Y.; Li, C.; Tang, J.; Luo, B. Quality-aware feature aggregation network for robust RGBT tracking. *IEEE Trans. Intell. Veh.* **2020**, *6*, 121–130. [CrossRef]
53. Liu, L.; Hua, Y.; Zhao, Q.; Huang, H.; Bovik, A.C. Blind image quality assessment by relative gradient statistics and adaboosting neural network. *Signal Process. Image Commun.* **2016**, *40*, 1–15. [CrossRef]
54. Liu, L.; Liu, B.; Huang, H.; Bovik, A.C. No-reference image quality assessment based on spatial and spectral entropies. *Signal Process. Image Commun.* **2014**, *29*, 856–863. [CrossRef]
55. Xue, W.; Mou, X.; Zhang, L.; Bovik, A.C.; Feng, X. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Trans. Image Process.* **2014**, *23*, 4850–4862. [CrossRef]
56. Crété-Roffet, F.; Dolmiere, T.; Ladret, P.; Nicolas, M. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In Proceedings of the SPIE Electronic Imaging Symposium Conference Human Vision and Electronic Imaging, San Jose, CA, USA, 12 February 2007; Volume 6492, pp. 196–206.

57. Palus, H. Colorfulness of the image: Definition, computation, and properties. In Proceedings of the Lightmetry and Light and Optics in Biomedicine 2004, SPIE, Warsaw, Poland, 20 April 2006; Volume 6158, pp. 42–47.
58. Yendrikhovskij, S.; Blommaert, F.J.; de Ridder, H. Optimizing color reproduction of natural images. In Proceedings of the Color and Imaging Conference. Society for Imaging Science and Technology, Scottsdale, AZ, USA, 17–20 November 1998; Volume 1998, pp. 140–145.
59. Hasler, D.; Suesstrunk, S.E. Measuring colorfulness in natural images. In Proceedings of the Human Vision and Electronic Imaging VIII, SPIE, Santa Clara, CA, USA, 17 June 2003; Volume 5007, pp. 87–95.
60. Berns, R.S. Extending CIELAB: Vividness, depth, and clarity. *Color Res. Appl.* **2014**, *39*, 322–330. [CrossRef]
61. Midtfjord, H.B.; Green, P.; Nussbaum, P. Vividness as a colour appearance attribute. In Proceedings of the Color and Imaging Conference. Society for Imaging Science and Technology, Washington, DC, USA, 16 June 2019; Volume 2019, pp. 308–313.
62. Chetverikov, D. Fundamental structural features in the visual world. In *Fundamental Structural Properties in Image and Pattern Analysis*; Citeseer: University Park, PA, USA, 1999.
63. Ou, L.C.; Luo, M.R.; Woodcock, A.; Wright, A. A study of colour emotion and colour preference. Part III: Colour preference modeling. *Color Res. Appl.* **2004**, *29*, 381–389. [CrossRef]
64. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06) IEEE, Hong Kong, China, 20–24 August 2006; Volume 3, pp. 850–855.
65. Hosang, J.; Benenson, R.; Schiele, B. Learning non-maximum suppression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4507–4515.
66. Campbell, F.W.; Robson, J.G. Application of Fourier analysis to the visibility of gratings. *J. Physiol.* **1968**, *197*, 551. [CrossRef]
67. Peli, E. Contrast in complex images. *JOSA A* **1990**, *7*, 2032–2040. [CrossRef]
68. Andre, T.; Antonini, M.; Barlaud, M.; Gray, R.M. Entropy-based distortion measure for image coding. In Proceedings of the IEEE 2006 International Conference on Image Processing, Atlanta, GA, USA, 8–11 October 2006; pp. 1157–1160.
69. Venkatanath, N.; Praneeth, D.; Bh, M.C.; Channappayya, S.S.; Medasani, S.S. Blind image quality evaluation using perception based features. In Proceedings of the IEEE 2015 Twenty First National Conference on Communications (NCC), Mumbai, India, 27 February–1 March 2015; pp. 1–6.
70. Ghosh, K.; Sarkar, S.; Bhaumik, K. A possible mechanism of zero-crossing detection using the concept of the extended classical receptive field of retinal ganglion cells. *Biol. Cybern.* **2005**, *93*, 1–5. [CrossRef] [PubMed]
71. Ghosh, K.; Sarkar, S.; Bhaumik, K. Understanding image structure from a new multi-scale representation of higher order derivative filters. *Image Vis. Comput.* **2007**, *25*, 1228–1238. [CrossRef]
72. Patil, S.B.; Patil, B. Automatic Detection of Microaneurysms in Retinal Fundus Images using Modified High Boost Filtering, Line Detectors and OC-SVM. In Proceedings of the IEEE 2020 International Conference on Industry 4.0 Technology (I4Tech), Pune, India, 13–15 February 2020; pp. 148–153.
73. Li, Q.; Lin, W.; Fang, Y. No-reference image quality assessment based on high order derivatives. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
74. Poynton, C.A. *A Technical Introduction to Digital Video*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1996.
75. 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]
76. Xu, L.; Lin, W.; Kuo, C.C.J. *Visual Quality Assessment by Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2015.
77. 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 Proceedings of the Visual Communications and Image Processing 2000, SPIE, Perth, Australia, 30 May 2000; Volume 4067, pp. 742–753.
78. 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.
79. Saad, M.A.; Bovik, A.C.; Charrier, C. Blind prediction of natural video quality. *IEEE Trans. Image Process.* **2014**, *23*, 1352–1365. [CrossRef] [PubMed]
80. Mittal, A.; Saad, M.A.; Bovik, A.C. A completely blind video integrity oracle. *IEEE Trans. Image Process.* **2015**, *25*, 289–300. [CrossRef]
81. 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]
82. Men, H.; Lin, H.; Saupe, D. Spatiotemporal feature combination model for no-reference video quality assessment. In Proceedings of the IEEE 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), Cagliari, Italy, 29 May–1 June 2018; pp. 1–3.
83. 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.
84. Tu, Z.; Wang, Y.; Birkbeck, N.; Adsumilli, B.; Bovik, A.C. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE Trans. Image Process.* **2021**, *30*, 4449–4464. [CrossRef]
85. Hosu, V.; Lin, H.; Sziranyi, T.; Saupe, D. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Trans. Image Process.* **2020**, *29*, 4041–4056. [CrossRef]

86. Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; Bovik, A. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3575–3585.
87. Ebenezer, J.P.; Shang, Z.; Wu, Y.; Wei, H.; Sethuraman, S.; Bovik, A.C. ChipQA: No-reference video quality prediction via space-time chips. *IEEE Trans. Image Process.* **2021**, *30*, 8059–8074. [CrossRef]