State-of-the-art image and video quality assessment with a metric based on an intrinsically nonlinear neural summation model

Raúl Luna[1*], Itziar Zabaleta[2], Marcelo Bertalmío[1]

[1]Institute of Optics, Spanish National Research Council (CSIC), Madrid, Spain

[2]Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain

*** Correspondence:**
Raúl Luna
raul.luna@csic.es

Keywords: visual perception, visual neuroscience, receptive field, INRF, computational modelling, image quality assessment, video quality assessment, high frame rate videos

Abstract

The development of automatic methods for image and video quality assessment that correlate well with the perception of human observers is a very challenging open problem in vision science, with numerous practical applications in disciplines such as image processing and computer vision, as well as in the media industry. In the past two decades, the goal of image quality research has been to improve upon classical metrics by developing models that emulate some aspects of the visual system, and while the progress has been considerable, state-of-the-art quality assessment methods still share a number of shortcomings, like their performance dropping considerably when they are tested on a database that is quite different from the one used to train them, or their significant limitations in predicting observer scores for high framerate videos. In this work we propose a novel objective method for image and video quality assessment that is based on the recently introduced Intrinsically Non-linear Receptive Field (INRF) formulation, a neural summation model that has been shown to be better at predicting neural activity and visual perception phenomena than the classical linear receptive field. Here we start by optimizing, on a classic image quality database, the four parameters of a very simple INRF-based metric, and proceed to test this metric on three other databases, showing that its performance equals or surpasses that of the state-of-the-art methods, some of them having millions of parameters. Next, we extend to the temporal domain this INRF image quality metric, and test it on a very recent video dataset of high definition, high-quality videos, of standard as well as high framerate, processed at different compression levels, and with thousands of human quality scores. Our results show that the proposed INRF-based video quality metric consistently outperforms, often by a wide margin, the state-of-the-art in video quality metrics.

1. INTRODUCTION

Image quality evaluation is of crucial importance in the media industry, where it has numerous practical applications, but also in applied disciplines such as image processing and computer vision, where it plays an important role in the development, optimization, and testing of algorithms. Subjective evaluation, consisting in measuring image quality by human beings, is costly and time-consuming. Therefore, the goal of objective quality assessment is to develop automatic methods that

produce quantitative measures that are consistent with the perception of human observers. But this is a very challenging open problem in vision science, given the limitations of current visual perception models and the way they are exacerbated by emerging image display technologies of ever-increasing resolution, contrast, color gamut and framerate (Bertalmío, 2019).

Image quality methods can be divided into three categories: full-reference methods, which compare an original image with a distorted version of it; reduced-reference methods, that compare some characteristics of the distorted and reference image since the complete reference image is not available; and no-reference methods (also called blind models), operating solely on the distorted image.

In this article we will focus on full-reference methods, which constitute the vast majority of image quality approaches. A simple solution, and possibly the most widely used metric to estimate image quality, is the peak signal-to-noise ratio (PSNR), which is a nonlinear transform of the mean square error (MSE) between the reference and the distorted images, another very popular metric. These metrics are simple to calculate, and they have a clear physical meaning; however, they are not very well correlated with perceived visual quality (Wang et al., 2004).

Therefore, in the last two decades, the goal of image quality assessment (IQA) research has been to improve these metrics by developing more sophisticated methods that mimic some aspects of the visual system. For instance, the Normalized Laplacian Pyramid Distance (NLPD) (Laparra et al., 2016) is based on transformations present in the early visual system: local luminance subtraction and local gain control, obtained from a decomposition of images using a Laplacian pyramid; the Structural Similarity Index (SSIM) (Wang et al., 2004) is based on the hypothesis that the human visual system is highly adapted for extracting structural information from the viewing field; the Feature Similarity Index (FSIM) (Zhang et al., 2011) is based on the assumption that the human visual system understands an image according to its low-level features, such as the phase congruency, which measures the significance of a local structure, and the image gradient magnitude, which encodes contrast information; and the Visual Information Fidelity Measure (VIF) (Sheik and Bovik, 2006) is based on natural scene statistics and models of the image degradation process and the human visual system. There are also learning-based methods that learn a metric from a set of training images and their corresponding perceptual scores. For instance, the Learned Perceptual Image Patch Similarity (LPIPS) metric (Zhang et al., 2018) is based on the hypothesis that perceptual similarity is a consequence of visual representations, as the authors found that internal activations of networks trained on high-level image classification tasks correspond well to human perceptual judgments; another example is the Deep Image Structure and Texture Similarity (DISTS) Metric (Ding et al., 2020), which uses a variant of the VGG convolutional neural network to construct a function that combines structure and texture similarity measurements between corresponding feature maps of the reference and distorted images; and PerceptNet (Hepburn et al., 2020), which is a convolutional neural network where the architecture reflects the structure and various stages in the human visual system: a cascade of canonical linear filters and divisive normalization layers simulate the retina-LGN-cortex pathway.

For video quality assessment (VQA), a simple option is to apply image quality metrics on a frame-by-frame basis, but this type of approach often provides a limited performance, especially in the case of high framerate (HFR) videos (Madhusudana et al., 2021). Therefore, the state-of-the-art in VQA are algorithms specifically developed for video, including the Spatio-temporal Reduced Reference Entropic Differences (ST-RRED) (Soundararajan and Bovik, 2012) or the Spatial Efficient Entropic Differencing for Quality Assessment (SpEED) (Bampis et al., 2017), both measuring quality

deviations by computing spatial and temporal entropic differences in the band-pass domain; Frame Rate dependent Quality Metric (FRQM) (Zhang et al., 2017), which outputs quality measurements by calculating absolute differences between sequences that have been temporally filtered by a wavelet; Video Multi-method Assessment Fusion (VMAF) (Li et al., 2016), which using a Support Vector Regressor fuses a frame-difference feature, a detail feature and features obtained from a Visual Information Fidelity (VIF) measure (Sheikh and Bovik, 2006); Deep Video Quality Assessor (deepVQA) (Kim et al., 2018), which combines a CNN model with a Convolutional Neural Aggregation Network (CNAN) used for temporal pooling; or the more recent Generalized Spatio-Temporal Index (GSTI) (Madhusudana et al., 2020), which calculates entropic differences between responses that have been temporally band-pass filtered.

Despite the notable advances in the field, it is important to point out that state-of-the-art quality assessment methods still share a number of shortcomings, like their performance dropping considerably when they are tested on a database that is quite different from the one used to train them (Ding et al., 2021), or their significant difficulties in predicting observer scores for HFR videos (Madhusudana et al., 2021). In this study we aim to overcome these limitations by estimating perceived image and video quality using a model for neural summation introduced recently, called the Intrinsically Non-linear Receptive Field (INRF) (Bertalmío et al., 2020) formulation; the INRF model successfully explains experimental data that linear receptive field models are unable to explain or do not explain accurately (Bertalmío et al., 2020), and it has been shown to be very promising as a tool to develop IQA methods given its ability to model complicated perceptual phenomena.

The contributions of this work are twofold. Firstly, we start by optimizing, on a classic image quality database, the four parameters of a very simple INRF-based metric, and proceed to test this metric on three other databases, showing that its performance equals or surpasses that of the state-of-the-art IQA methods, some of them having millions of parameters. Secondly, we extend to the temporal domain this INRF image quality metric, and test it on a very recent video dataset of high definition, high-quality videos, of standard as well as high framerate, processed at different compression levels, and with thousands of human quality scores. Our results show that the proposed INRF-based video quality metric consistently outperforms, often by a wide margin, the state-of-the-art in VQA.

The structure of this manuscript is as follows. Section 2 explains the INRF model and how it is used for IQA and VQA. Section 3 provides details on how the INRF parameters are optimized for IQA and VQA usage, describes the different datasets on which performance is tested and explains further points on how the assessment of INRF for IQA and VQA models is performed. Section 4 shows the performance of INRF, employed for IQA and VQA, on different datasets, and results are compared with performance of other state-of-the-art IQA and VQA algorithms on those same datasets. Finally, in Section 5, results and their implications are discussed.

2. PROPOSED METHODS FOR IQA AND VQA

2.1. Overview of the INRF neural summation model

In vision science, the receptive field (RF) of a neuron is the extent of the visual field where light influences the neuron's response. In the "standard model" of vision, the first stage is a filtering operation consisting of multiplying the intensities at each local region of an image stimulus by the values of a filter (the weights of the RF), and summing the resulting intensities (Carandini et al., 2005); this weighted sum may then be normalized by the responses of neighboring neurons and passed through a point-wise nonlinearity. Many scientists have come to accept this linear-plus-nonlinear (L+NL) formulation as a working model of the visual system (Olshausen and Field, 2005), both in visual neuroscience and in visual perception, and while there have been considerable improvements on, and extensions to, the standard model, the linear RF remains as the foundation of most vision models. But there are a number of problems that are inherent to considering the RF as having a linear form, of which we will highlight three:

- adaptation makes the linear RF change with the input and, in fact, the linear RF has been observed to have different sizes, orientations, preferred directions or even different polarity (ON/OFF) for different stimuli (Cavanaugh et al., 2002; Coen-Cagli et al., 2012; Jansen et al., 2018);
- a linear RF depends on the choice of basis functions used to estimate it (Vilankar and Field, 2017);
- the linear RF is not supported by more recent neuroscience, and a growing number of neuroscience studies show that in general individual neurons *cannot* be modeled as a linear RF followed by an output nonlinearity (Poirazi et al., 2003; Polsky et al., 2004; London and Häuser, 2005; Silver, 2010; Rodrigues et al., 2021).

In contrast, the INRF formulation is a physiologically-plausible single-neuron summation model which, unlike the linear RF:

- embodies the efficient representation principle and can remain constant in situations where the linear RF must change with the input;
- is a generalization of the linear RF that is much more powerful in representing nonlinear functions;
- is consistent with more recent studies on dendritic computations.

The INRF equation for the response of a single neuron at location x is:

$$INRF(x) = \sum_i m_i\, I(y_i) - \lambda \sum_i w_i\, \sigma\left(I(y_i) - \sum_j g\left(y_j - x\right) I\left(y_j\right) \right) \quad (1)$$

where m_i stands for a 2D kernel $m(x, y_i)$, locations y_i are neighbors of x, λ is a scalar, w_i stands for a 2D kernel $w(x, y_i)$, σ represents a non-linearity, and g is also a 2D kernel.

The model is based on knowledge about dendritic cells: some dendritic branches act as non-linear units, a single non-linearity is not enough to model dendritic computations, and there is feedback from the neuron soma to the dendrites. In the INRF model, some dendrites are linear and their contributions are summed with weights m_i, and some other dendrites are nonlinear and their contributions are summed with weights w_i. The feedback from the soma is reflected in the shifting term of the nonlinearity σ, expressed by the term $\sum_j g\left(y_j - x\right)I\left(y_j\right)$, which has the effect of making the nonlinearity change for each contributing location y_i.

Using a single, fixed INRF module where the kernels m, w, g have Gaussian form and the nonlinearity is a power-law sigmoid, and applying it to grayscale images, where x and y_i now denote pixel locations and $I(\cdot)$ are pixel values, the model response $INRF(\cdot)$ emulates the perceived image and can explain several visual perception phenomena that challenge the standard model:

- The "crispening" effect. Brightness perception curves show "crispening" (slope increase) at the surround luminance level when the background is uniform, but the phenomenon disappears for salt and pepper background (Kane and Bertalmío, 2019). The INRF response qualitatively replicates both cases with a fixed set of parameters, which is not possible with a linear RF formulation.

- White's illusion under noise. The INRF output qualitatively predicts the observers' response to White's illusion when bandpass noise is added, when in Betz et al. (2015) none of the vision models that were tried, based on linear RFs, were able to replicate this behavior.

- Light/dark asymmetry (the *"irradiation illusion"*). This phenomenon can be reproduced with a fixed INRF formulation, while a L+NL model needs to change with the stimulus (Kremkow et al., 2014).

- Second-order masking. In Henning et al. (1975), Henning et al. show evidence that contradicts the basic tenet of the standard model for visual perception, the hypothesis that the visual system analyses spatial patterns in independent narrowly-tuned bands of spatial frequency. The INRF output, on the other hand, is able to explain those experimental results without any problem.

In short, we could say that the INRF formulation, a non-linear transform of light intensity values, is a good estimator of brightness perception. This is a very valuable property when creating an image quality assessment method, as we discuss next.

2.2. IQA with the INRF model

Given that applying the INRF model to a grayscale image produces a result that closely resembles how the image is perceived, Bertalmío et al. (2020) proposed a very simple image quality metric: given an image I, and its distorted version I_D, the INRF transformation is applied to both of the images, obtaining O and O_D, and then the root mean square error (RMSE) between the processed images is computed. The underlying idea here was that the INRF model performs a sort of "perceptual linearization", a nonlinear transform that brings *light-intensity* images (whose direct comparison, with metrics like RMSE or PSNR, does not correlate well with perception), to the space corresponding to *perceptual* images (which can now be compared with a simple Euclidean metric

like RMSE). This metric had only five parameters that were optimized for the "crispening" brightness perception experiment mentioned above (with σ a power-law sigmoid and g a delta function) on a handful of synthetic images, and despite this fact and the simplicity of the metric, when tested on the natural image database TID2013 (Ponomarenko et al., 2015) it was shown to have a performance very similar to that of the state-of-the-art deep learning perceptual metric LPIPS (Zhang et al., 2018), with over 24 million parameters and close to 500K human judgements for labeling pair-comparison preferences on the 160,000+ natural images it used for training.

Based on this very promising result, here we propose a full-reference INRF-based IQA method in the following way:

1 Given a grayscale image I, the INRF transformation applied to it produces an image O whose value at each pixel location x is computed as:

$$O(x) = INRF(x) = m \star I(x) - \lambda \sum_i w_i \, S\big(I(y_i) - g \star I(x)\big) \quad (2)$$

where the kernels m, w, g are 2D Gaussians of standard deviations $\sigma_m, \sigma_w, \sigma_g$ respectively, λ is a scalar, S is a sigmoid that has the form of an **atan** function, and the symbol $\star$ denotes the 2D convolution.

2 The INRF-IQA value comparing image I and its distorted version I_D is computed as:

$$INRF\text{-}IQA(I, I_D) = \sqrt{MSE(O, O_D)} = \sqrt{MSE\big(INRF(I), INRF(I_D)\big)} \quad (3)$$

where the INRF transform of an image is computed as in Eq. 2.

3 The values for the four parameters of the metric, namely $\sigma_m, \sigma_w, \sigma_g$ and λ, are chosen so as to maximize the correlation between the INRF-IQA perceptual distance of Eq. 3 and the mean opinion scores (MOS) of human observers over the large scale natural image database TID2008 (Ponomarenko et al., 2009) (see Methods).

2.3. VQA with the INRF model

We propose as a full-reference INRF-based VQA metric a straightforward combination of the INRF-IQA metric of Eq. 3 with a simple temporal pooling strategy: the image metric is applied frame-by-frame, and then the results are averaged. That is, given a reference video V and a distorted version V_D, the output of our proposed INRF-VQA metric is:

$$INRF\text{-}VQA(V, V_D) = mean\big\{INRF\text{-}IQA_s\big(I^i, {I^i}_D\big)\big\} \quad (4)$$

where I^i and ${I^i}_D$ are, respectively, the i-th frame of V and V_D, and the IQA metric $INRF\text{-}IQA_s$ has the same form as in Eq. 3 and the same value of 3 for the parameter λ, but the spatial kernel sizes $\sigma_m, \sigma_w, \sigma_g$ are scaled by a factor f that denotes the ratio between the size of the frames in the video V and the size of the images in TID2008 (that were used to optimize $\sigma_m, \sigma_w, \sigma_g$). For instance, if V is a

2K video of resolution $1920x1080$, and given that images in TID2008 are of size $512x384$, the scaling factor is $f = \frac{1920}{512} = 3.75$.

3. METHODS

3.1. Optimization details

The parameter values of INRF-IQA (Eq. 3) are found through optimization, and then, they are also used for INRF-VQA (Eq. 4) with proper rescaling as explained in section "2.3. VQA with the INRF model". Optimization is performed on the TID2008 dataset (Ponomarenko et al., 2009) to maximize PLCC between existing opinion scores of observers for the images in TID2008 and INRF-IQA scores. In particular, a grid search optimization method is used to optimize the INRF-IQA parameter values $\sigma_m, \sigma_w, \sigma_g$ and λ. A grid of parameter values defines the search space, with every position in the grid evaluated. The optimal parameters are those generating the highest correlation. The grid values are: for σ_m, $\{1.74\}$, for σ_w, $\{10,25,50,150\}$, for σ_g, $\{1\}$, and for λ, $\{1.5,3,4.5\}$. The optimal parameters found are $\sigma_m = 1.74, \sigma_w = 25, \sigma_g = 1$ and $\lambda = 3$.

3.2. IQA datasets

We use the LIVE (Sheikh et al., 2006), CSIQ and TID2013 (Ponomarenko et al., 2015) datasets to test the performance of our INRF-IQA metric on images; each dataset also containing subjective image quality scores from observers for those same images. The LIVE dataset consists of 779 images. It has a total of 29 reference images whose distorted versions are achieved by applying 5 different types of distortion (Gaussian blur, additive white Gaussian noise, JPEG compression, JP2K compression and a simulated fast fading Rayleigh channel) with different distortion levels. The CISQ dataset has 866 images, with 30 being reference ones. Their distorted versions are obtained through Gaussian blur, Gaussian pink noise, Gaussian white noise, JPEG compression, JP2K compression and contrast change. The TID2013 dataset extends the previous TID2008 one (Ponomarenko et al., 2009). It is composed of 3000 images, out of which 25 are reference ones. These are distorted to achieve the rest of the images by applying 24 distortion types (7 new types of distortions with respect to TID2008) each with 5 distortion levels. Distortion types are rich, spanning from Gaussian noise, Gaussian blur, lossy compression of noisy images, and distortions such as JPEG to more uncommon distortions like non-eccentricity pattern noise.

3.3. VQA dataset

We use the LIVE-YouTube-HFR dataset (Madhusudana et al., 2021), available online for public use, to test the performance of our INRF-VQA model on HFR videos. This dataset contains videos for which subjective measurements of video quality exist (See Madhusudana et al., 2021, their section "II. LIVE-YouTube-HFR DATABASE" for a detailed description of the dataset). It spans 16 different video contents. That is, 16 possible categories of videos, each showing a different progressively scanned natural scene (See Madhusudana et al., 2021, their Figure 1 to see the different video contents). Out of those 16 contents, 11 have 2K spatial resolution and 5 have 4K resolution. Within each of the video contents, videos with different frame rates exist, out of which we use 120 fps, 60 fps and 30 fps. Each video content with a given frame rate has 5 possible compression levels (FFmpeg VP9 compression [Mukherjee et al., 2015] and single-pass encoding varying the Constant

Rate Factor [CRF]). For instance, a given content and its videos with a given frame rate sum a total of 5 videos: 1 video with lossless compression (CRF=0) + 4 videos with compression levels ranging from CRF=4 to CRF=63. Finally, for each video content, the 120 fps video with lossless compression (CRF=0) is referred to as the reference sequence. All the remaining videos within a content; that is, 120 fps videos with CRF values larger than 0, and 60 fps and 30 fps videos, are the distorted sequences.

3.4. Evaluation of INRF-IQA and INRF-VQA models

Performance of INRF-IQA is evaluated on the standardized image quality datasets LIVE (Sheikh et al., 2006), CSIQ and TID2013 (Ponomarenko et al., 2013; Ponomarenko et al., 2015), which consist of reference and artificially distorted images. On the other hand, we assess performance of INRF-VQA on the LIVE-YT-HFR dataset (Madhusudana et al., 2021), which contains undistorted reference videos and distorted versions of them made by lowering frame rate and introducing compression artifacts. Even though these videos have durations ranging from 6 to 10 seconds, we only compute INRF-VQA over the last 2 seconds of the videos to speed performance. Existing subjective image quality and video quality scores in each of the datasets are respectively correlated with our INRF-IQA and INRF-VQA metrics. In the case of INRF-IQA, SRCC is calculated. For INRF-VQA, SRCC, Kendal's rank correlation coefficient (KRCC) and PLCC are obtained. The same as in Antkowiak et al. (2000) or Madhusudana et al. (2021), before calculating PLCC, predicted objective video quality scores are passed through a four-parameter sigmoid function defined as:

$$\hat{y}(x) = \beta_2 \frac{\beta_1 - \beta_2}{1 + e^{-\left(\frac{x-\beta_3}{|\beta_4|}\right)}} \quad (5)$$

where x stands for the raw INRF-VQA scores, andβ_1, β_2, β_3 and β_4 are its parameters. These are optimized to maximize PLCC between $\hat{y}$ and the existing subjective video quality scores using the Matlab function "fminsearch".

4. RESULTS AND COMPARISONS

4.1. IQA

Performance of our INRF-IQA metric is evaluated on images from the LIVE (Sheikh et al., 2006), CSIQ and TID2013 (Ponomarenko et al., 2015) datasets, for which subjective image quality scores from observers also exist. Each dataset consists of a number of reference images out of which distorted images are created by applying several distortion types (e.g. Gaussian blur, JP2K compression, etc.) at different distortion levels.

To evaluate objective IQA performance, the quality score of a given IQA metric for each image in the datasets needs to be correlated with existing subjective image quality scores for the same images (see Methods). Table 1 shows a comparison in the performance of different full-reference IQA metrics against our INRF-IQA metric. Spearman rank correlation coefficients (SRCC) are shown for each, evaluated on the the LIVE, CSIQ and TID2013 datasets.

Table 1. Numbers indicate Spearman rank correlation coefficients (SRCC). The INRF-IQA metric is compared against a set of full-reference image quality methods: MS-SSIM (Wang et al., 2003), CW-SSIM (Wang and Simoncelli, 2005), VIF (Sheikh and Bovik, 2006), NLPD (Laparra et al., 2016), GMSD (Xue et al., 2014), MAD (Larson and Chandler, 2010), FSIM (Zhang et al., 2011), VSI (Zhang et al., 2014), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2020), and PerceptNet (Hepburn et al., 2020). The best three correlation values per column are marked in bold. Adapted table from Ding et al. (2021).

	LIVE	CSIQ	TID2013	(mean)
MS-SSIM	0,951	0,886	0,782	(0,873)
CW-SSIM	0,781	0,738	0,680	(0,733)
VIF	**0,963**	0,911	0,676	(0,850)
NLPD	0,938	0,937	0,800	(0,892)
GMSD	0,960	**0,950**	**0,804**	**(0,905)**
MAD	0,960	**0,941**	0,773	(0,891)
FSIM	**0,963**	0,916	**0,802**	(0,894)
VSI	0,950	0,923	0,793	(0,889)
LPIPS	0,932	0,837	0,616	(0,795)
DISTS	0,942	0,905	0,764	(0,870)
PerceptNet	**0,98**	**0,96**	**0,87**	**(0,93)**
INRF-IQA	0,947	0,952	**0,802**	**(0,900)**

Results evidence a good performance of INRF-IQA in comparison to CNN-based models such as LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020), as well as other extensively used classical methods like NLPD (Laparra et al., 2016). Overall, the best IQA performance is observed for PerceptNet (Hepburn et al., 2020), GMSD (Xue et al., 2014), and our INRF-IQA metric.

4.2. VQA

Performance of our INRF-VQA metric is evaluated on videos from the LIVE-YouTube-HFR dataset (Madhusudana et al., 2021). This dataset consists of reference videos with 120 fps frame rate for which distorted versions are generated by reducing their frame rate (e.g. 60 fps, 30 fps and more) and applying different compression levels. Subjective measurements of video quality are provided for each of the videos.

To assess the performance of a VQA algorithm, its VQA objective metric for each video needs to be correlated with existing subjective video quality scores. In this case, final subjective video quality scores to employ during correlations are obtained by subtracting the quality score of a distorted video from the quality score of its respective reference video.

It is important to note that our INRF-VQA metric, as well as many other full-reference metrics, needs reference and distorted video sequences to have the same number of frames. For this reason, when a distorted video has a lower frame rate than the reference video, either the reference video must be downsampled to match the number of frames in the distorted video or the distorted video must be upsampled to match the reference. Madhusudana et al., 2021 (see their Table 5), who test the performance of several VQA algorithms on the LIVE-YouTube-HFR dataset, use naive temporal upsampling by frame duplication of distorted videos. In doing so, they argue that downsampling may introduce undesired temporal artifacts in reference videos. VMAF (Li et al., 2016) is a very successful, state-of-the-art, full-reference VQA metric proposed and used by Netflix, and its performance comparison against our INRF-VQA metric can be seen in Table 2. Performance is shown for different frame rates in the LIVE-YT-HFR dataset: 120 fps, 60 fps and 30 fps; and in the case of 60 fps and 30 fps, both an upsampling (results taken from Madhusudana et al., 2021) and a downsampling approach are used. Upsampling is achieved by frame duplication of distorted videos and downsampling is done through frame dropping of reference videos. This way, reference and distorted videos have the same number of frames. Spearman and Pearson correlation coefficients (SRCC and PLCC) of the objective VQA metrics with subjective video quality scores are displayed.

Table 2. Numbers indicate SRCC and PLCC for frame rates of 120 fps, 60 fps and 30 fps. The INRF-VQA metric is compared against VMAF (Li et al., 2016). For 60 fps and 30 fps, INRF-VQA and VMAF results are shown both using frame duplication of distorted videos (INRF-VQA$_{dup}$ and VMAF$_{dup}$) and frame dropping of reference videos (INRF-VQA$_{drop}$ and VMAF$_{drop}$). VMAF$_{dup}$ results are taken from Madhusudana et al. (2021), their Table 5.

	120 fps		**60 fps**		**30 fps**	
	SRCC	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**
VMAF$_{dup}$	0.7943	0.7844	0.5408	0.6015	0.2855	0.3740
VMAF$_{drop}$			0.6664	0.8003	0.5981	0.8713
INRF-VQA$_{dup}$	0.8452	0.8169	0.2945	0.3639	0.1003	0.2440
INRF-VQA$_{drop}$			0.7515	0.8193	0.6292	0.8715

For 60 fps and 30 fps, and both for INRF-VQA and VMAF, performance is seen to improve when a frame dropping strategy rather than an upsampling one is used (for which correlations are low). Because of this, together with the low computational cost of frame dropping, we use this approach as our preferred choice to evaluate INRF-VQA for 30 fps and 60 fps.

Table 3 shows a performance comparison of our INRF-VQA metric (using frame dropping when 60 fps and 30 fps videos are evaluated) against other state-of-the-art full-reference VQA metrics (see Methods). SRCC and PLCC results of the objective VQA metrics with subjective video quality scores are shown for different frame rates in the LIVE-YT-HFR dataset.

The results in Table 3 demonstrate that INRF-VQA consistently outperforms all state-of-the-art algorithms for all frame rates tested, including the highest frame rate of 120 fps (for a fair comparison, results for VMAF for 60 fps and 30 fps were also computed using frame dropping, and we do not rule out that other methods could also benefit from carrying out the validation in this manner).

Table 3. Numbers indicate SRCC and PLCC for frame rates of 120 fps, 60 fps and 30 fps. The INRF-VQA metric is compared against a set of full-reference video quality methods (See Madhusudana et al., 2021, their Table 5): PSNR, SSIM (Wang et al., 2003), MS-SSIM (Wang et al., 2004), FSIM (Zhang et al., 2011), ST-RRED (Soundararajan and Bovik, 2012), SpEED (Bampis et al., 2017), FRQM (Zhang et al., 2017), VMAF (Li et al., 2016), deepVQA (Kim et al., 2018) and GSTI (Madhusudana et al., 2020). For 60 fps and 30 fps, results for all VQA algorithms are shown using naive temporal upsampling of distorted videos except for INRF-VQA and VMAF, where frame dropping of reference videos is used. The best three correlation values per column are marked in bold. Adapted from Madhusudana et al. (2021).

	120 fps		60 fps		30 fps	
	SRCC	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**
PSNR	0.6019	0.5937	**0.6202**	0.5719	0.4414	0.4179
SSIM	**0.7485**	0.6726	0.2123	0.1845	0.1108	0.0816
M-SSIM	0.6165	0.5843	0.2516	0.1900	0.1929	0.1112
FSIM	0.7053	0.6368	0.3450	0.2776	0.2487	0.1786
ST-RRED	0.6745	0.5906	0.5062	0.4457	0.1188	0.0307
SpEED	0.6827	0.6097	0.1824	0.1110	0.2278	0.0896
FRQM	-	-	0.0947	0.0309	0.0983	0.0854
VMAF	**0.7943**	**0.7844**	**0.6664**	**0.8003**	**0.5981**	**0.8713**
deepVQA	0.6865	0.6209	0.2527	0.1652	0.1353	0.1059
GSTI	0.7390	**0.7003**	0.6015	**0.7566**	**0.4758**	**0.6689**
INRF-VQA	**0.8452**	**0.8169**	**0.7515**	**0.8193**	**0.6292**	**0.8715**

5. DISCUSSION

In this study we have taken a recent neural summation model and used it as a foundation for novel metrics for image and video quality assessment.

Our validation, on popular datasets of observer scores, shows that our proposed metric for IQA compares very well with the state-of-the-art and, very importantly, that its performance is consistently very good and does not drop for different datasets, unlike what many methods are prone to do (especially those based on deep learning techniques).

Our metric for VQA is shown to outperform all state-of-the-art models, often by a wide margin, for all framerates considered, including a high frame rate of 120fps. Our proposed INRF-VQA metric is a straightforward extension of the INRF-IQA method, in which metric values are computed on a frame-by-frame basis and then averaged over time to produce a single score for the video: the fact that this simple temporal extension of an IQA method works so remarkably well for VQA challenges a common assumption in the literature, where it is thought that the best VQA metrics must be developed specifically for video. Contrary to our findings, Madhusudana et al. (2021) claim that applying IQA models on a frame-by-frame basis and temporally pooling the frame-by-frame outcome often yields limited performance because the temporal information present in videos is not used effectively.

Arguably, the distortions caused by the changes in frame rate are much less perceptually relevant than the artifacts created by compression. This would explain why the full-reference metrics not specifically designed to work with different frame rates (such as VMAF and INRF-VQA) correlate well with the observers' responses even when the information about the temporal differences is reduced (i.e. when reference frames are dropped).

For INRF-VQA we have tested other temporal pooling strategies besides the mean average, as well as modifying the INRF-IQA computation for each frame so that it takes into account the values of

previous frames: in both cases, the results do not improve upon what we can achieve with the original, simple formulation for INRF-VQA (see Supplementary Material). However, we do not rule out that other, more elaborate formulations could prove better (e.g. embedding the temporal processing into the INRF formulation in a non-separable way, as opposed to what we currently do which is spatial processing followed by temporal processing).

An interesting observation produced by our testing is that, when validating VQA methods that are applied to videos of different framerate, each of the two options (frame dropping and frame duplication) is better suited for different types of metrics. For instance, metrics that are very sensitive to small displacements of image points, like VMAF or INRF-VQA, show a much better correlation to observer scores when they are evaluated using frame dropping instead of frame duplication, which makes sense precisely because frame duplication introduces many small displacements in the image pair to be compared (since these images will correspond to different times, in most cases). For example, when using frame duplication for evaluation (Madhusudana et al., 2021), the method GSTI (Madhusudana et al., 2020) greatly outperforms VMAF, but the reverse is true if we use frame duplication just for GSTI and frame dropping for VMAF.

Future work may test INRF on new datasets. In this regard, we are currently working on extending both INRF-IQA and INRF-VQA so that they consider color and can also be applied to high dynamic range (HDR) and wide color gamut (WCG) as well as 8K imagery. Also, we are working on a spatio-temporal formulation of INRF-VQA.

6. REFERENCES

Bampis, C. G., Gupta, P., Soundararajan, R., & Bovik, A. C. (2017). SpEED-QA: Spatial efficient entropic differencing for image and video quality. *IEEE signal processing letters, 24*(9), 1333-1337.

Bertalmío, M. (2019). *Vision Models for High Dynamic Range and Wide Colour Gamut Imaging: Techniques and Applications* (Academic Press)

Bertalmío, M., Gomez-Villa, A., Mart´ın, A., Vazquez-Corral, J., Kane, D., and Malo, J. (2020). Evidence for the intrinsically nonlinear nature of receptive fields in vision. *Scientific reports* 10, 1–15. Nature Publishing Group

Betz, T., Shapley, R., Wichmann, F. A., and Maertens, M. (2015). Testing the role of luminance edges in White's illusion with contour adaptation. *Journal of vision* 15, 14–14

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., et al. (2005). Do we know what the early visual system does? *Journal of Neuroscience* 25, 10577–10597

Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002). Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of neurophysiology* 88, 2547–2556

Coen-Cagli, R., Dayan, P., and Schwartz, O. (2012). Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS computational biology* 8, e1002405

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. *CoRR* abs/2004.07728

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2021). Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 244, 1258–1281

Fukushima, Y., Hara, K., & Kimura, M. (1986). A spatio-temporal model of ganglion cell receptive field in the cat retina. *Biological cybernetics*, *54*(2), 91-98.

Henning, G. B., Hertz, B. G., and Broadbent, D. (1975). Some experiments bearing on the hypothesis that the visual system analyses spatial patterns in independent bands of spatial frequency. *Vision research* 15, 247, 887–897

Hepburn, A., Laparra, V., Malo, J., McConville, R., and Santos-Rodriguez, R. (2020). Perceptnet: A human visual system inspired neural network for estimating perceptual distance. In *2020 IEEE International Conference on Image Processing (ICIP)* (IEEE), 121–125

Jansen, M., Jin, J., Li, X., Lashgari, R., Kremkow, J., Bereshpolova, Y., et al. (2018). Cortical balance between on and off visual responses is modulated by the spatial properties of the visual stimulus. *Cerebral Cortex* 29, 336–355

Kane, D. and Bertalmío, M. (2019). A reevaluation of Whittle (1986, 1992) reveals the link between detection thresholds, discrimination thresholds, and brightness perception. *Journal of vision* 19, 16–16

Kaplan, E., & Benardete, E. (2001). The dynamics of primate retinal ganglion cells. *Progress in brain research*, 134, 17-34.

Kim, W., Kim, J., Ahn, S., Kim, J., & Lee, S. (2018). Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. *In Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 219-234).

Kremkow, J., Jin, J., Komban, S. J., Wang, Y., Lashgari, R., Li, X., et al. (2014). Neuronal nonlinearity explains greater visual spatial resolution for darks than lights. *Proceedings of the National Academy of Sciences*, 201310442

Laparra, V., Ballé, J., Berardino, A., and Simoncelli, E. (2016). Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging* 2016, 1–6. doi:10.2352/ISSN.2470-1173.2016.16. HVEI-103

Larson, E. C. and Chandler, D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging* 19, 011006

Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A., & Manohara, M. (2016). Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, *6*(2).

London, M. and Häusser, M. (2005). Dendritic computation. *Annu. Rev. Neurosci.* 28, 503–532

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., & Bovik, A. C. (2020). Capturing video frame rate variations via entropic differencing. *IEEE Signal Processing Letters*, *27*, 1809-1813.

Madhusudana, P. C., Yu, X., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. (2021). Subjective and objective quality assessment of high frame rate videos. *IEEE Access* 9, 108069–108082

Mukherjee, D., Han, J., Bankoski, J., Bultje, R., Grange, A., Koleszar, J., Wilkins, P., & Xu, Y. (2015). A technical overview of vp9—the latest open-source video codec. *SMPTE Motion Imaging Journal*, *124*(1), 44-54.

Olshausen, B. A. and Field, D. J. (2005). How close are we to understanding V1? *Neural computation* 17, 1665–1699

Poirazi, P., Brannon, T., and Mel, B. W. (2003). Pyramidal neuron as two-layer neural network. *Neuron* 37, 989–999

Polsky, A., Mel, B. W., and Schiller, J. (2004). Computational subunits in thin dendrites of pyramidal cells. *Nature neuroscience* 7, 621

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., et al. (2015). Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* 30, 57–77

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F. (2009). TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. *Advances of Modern Radioelectronics* 10, 30–45

Reid, R. C., & Shapley, R. M. (2002). Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus. *Journal of Neuroscience*, *22*(14), 6158-6175.

Rodrigues, Y. E., Tigaret, C. M., Marie, H., O'Donnell, C., and Veltz, R. (2021). A stochastic model of hippocampal synaptic plasticity with geometrical readout of enzyme dynamics. *bioRxiv*

Seshadrinathan, K., & Bovik, A. C. (2007). A structural similarity metric for video based on motion models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (Vol. 1, pp. I-869). IEEE.

Seshadrinathan, K., & Bovik, A. C. (2009). Motion tuned spatio-temporal quality assessment of natural videos. *IEEE transactions on image processing*, *19*(2), 335-350.

Sheikh, H. and Bovik, A. (2006). Image information and visual quality. *IEEE Transactions on Image Processing* 15, 430–444. doi:10.1109/TIP.2005.859378

Sheikh, H. R., Sabir, M. F., & Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, *15*(11), 3440-3451.

Soundararajan, R., & Bovik, A. C. (2012). Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(4), 684-694.

Silver, R. A. (2010). Neuronal arithmetic. *Nature Reviews Neuroscience* 11, 474–489

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, *12*(3), 289.

Vilankar, K. P. and Field, D. J. (2017). Selectivity, hyperselectivity, and the tuning of V1 neurons. *Journal of vision* 17, 9–9

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 600–612. doi:10.1109/TIP.2003. 819861

Wang, Z., Simoncelli, E., and Bovik, A. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* (IEEE), vol. 2, 1398–1402

Wang, Z. and Simoncelli, E. P. (2005). Translation insensitive image similarity in complex wavelet domain. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE), vol. 2, ii–573

Wang, Z., & Li, Q. (2007). Video quality assessment using a statistical model of human visual speed perception. *JOSA A*, *24*(12), B61-B69.

Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2014). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing* 23, 684–695. doi:10.1109/TIP.2013.2293423

Zhang, F., Mackin, A., & Bull, D. R. (2017). A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling. *In 2017 IEEE International Conference on Image Processing (ICIP)* (pp. 300-304). IEEE.

Zhang, L., Shen, Y., and Li, H. (2014). VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing* 23, 4270–4281. doi:10.1109/TIP.2014.2346028

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing* 20, 2378–2386 doi:10.1109/TIP.2011.2109730

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*

7. Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8. Author Contributions

RL and MB wrote the manuscript. RL did the final version of the INRF-VQA code and carried out the INRF-VQA experiments. ZL did the INRF-IQA version of the code and carried out the INRF-IQA experiments. MB conceptualized the original idea. All authors reviewed the manuscript.

9. Funding

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 952027 (project AdMiRe). RL is supported by a Juan de la Cierva-Formación fellowship (FJC2020-044084-I) funded by MCIN/AEI /10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

10. Acknowledgments

The authors would like to thank Adrián Martín for providing us with a fast implementation of the INRF code; Jesús Malo for providing the TID2008 and TID2013 datasets; Pavan C. Madhusudana, for facilitating data from the LIVE-YouTube-HFR dataset and for his explanations about it; and Yasuko Sugito for her helpful suggestions on running VMAF.

Supplementary Material

1. Temporal pooling methods in INRF-VQA metrics

INRF-VQA metrics provide a unique video quality score out of temporaly pooling a set of frame-wise INRF-IQA scores. A simple temporal pooling strategy is to compute the arithmetic mean from frame-by-frame scores, as seen in Eq. 4 from the manuscript. We rewrite that equation here for convenience, renaming it as Eq. 1:

$$\text{INRF-VQA}_{mean}(V, V_D) = \text{mean}\{\text{INRF-IQA}_s(I^i, I^i{}_D)\} \qquad (1)$$

where I^i and $I^i{}_D$ are, respectively, the i-th frame of V and V_D (a reference video and its distorted version respectively), and $INRF\text{-}IQA_s$ is the IQA metric, defined as in Eq. 3 from the manuscript.

Other possible temporal pooling strategies are computing the median, the maximum or the minimum instead of the mean, also Minkowski summation, kth percentile or averaging only the information in the last frames of the videos (we refer to this as last seconds pooling method). We define these in the following lines.

Median

$$\text{INRF-VQA}_{median}(V, V_D) = \text{median}\{\text{INRF-IQA}_s(I^i, I^i{}_D)\} \qquad (2)$$

Maximum

$$\text{INRF-VQA}_{max}(V, V_D) = \max\{\text{INRF-IQA}_s(I^i, I^i{}_D)\} \qquad (3)$$

Minimum

$$\text{INRF-VQA}_{min}(V, V_D) = \min\{\text{INRF-IQA}_s(I^i, I^i{}_D)\} \qquad (4)$$

Considering that large frame-wise INRF-IQA scores (before being temporally pooled) imply a large difference between a distorted and the reference video sequence, frames with such high values are perceived with poor quality. To understand this better, the same thing can be thought the other way round: when the difference in frame-wise INRF-IQA scores between the reference and a distorted video is small, this means that the distorted video has a quality similar to that of the undistorted reference, and therefore, perceived quality is good. Thus, pooling methods like taking the maximum

or minimum stress the influence of either the frame with the worst quality or the frame with the best one.

Minkowski summation

In this same line, the Minkowski summation method emphasizes the influence of poor quality frames as its exponent p, given in the following formula, increases:

$$\text{INRF-VQA}_{Mink}(V, V_D) = \left[\frac{1}{F}\sum_{f=1}^{F} \text{INRF-IQA}_s\left(I^i, I^i{}_D\right)\right]^{\frac{1}{p}} \qquad (5)$$

where V and V_D are reference and distorted video sequences respectively, I^i and $I^i{}_D$ are, respectively, the i-th frame of V and V_D (a reference video and its distorted version respectively) and $INRF\text{-}IQA_s$ is the IQA metric, defined as in Eq. 3 from the manuscript. The total number of frames is F. When p has a value of 1, calculating Minkowski summation is analogous to calculating the arithmetic mean. With a value of -1, it is analogous to calculating the harmonic mean, and a value of -2 equals calculating the quadratic mean.

k^{th} percentile

k^{th} percentile is another temporal pooling method that also emphasizes the effect of the poorest quality frames by computing the arithmetic mean using only frame-wise INRF-IQA scores (before being temporally pooled) above a certain percentile:

$$INRF\text{-}VQA_{kth}(V, V_D) = \frac{1}{|P_{\uparrow k\%}|}\sum_{f=1}^{F} \text{INRF-IQA}_s\left(I^i, I^i{}_D\right) \qquad (6)$$

where $|P_{\uparrow k\%}|$ stands for the set of highest $k\%$ scores (i.e. poorer quality frames).

Last seconds

Finally, last seconds pooling method calculates the mean quality score only by considering the most recent seconds of a video. Seconds need to be translated to “number of frames” consistently with the frame rate of the video being analyzed (e.g. the last two seconds of a 60 fps video correspond to its last $2\ s \times 60\ Hz = 120$ frames):

$$INRF\text{-}VQA_{last}(V, V_D) = \frac{1}{F}\sum_{t=T-F+1}^{T} \text{INRF-IQA}_s\left(I^i, I^i{}_D\right) \qquad (7)$$

where F denotes the number of most recent frames and T is the total number of frames. Considering that the whole duration of a video were considered for analysis, this method simulates a recency effect when evaluating video quality, with high values of F implying a weaker effect. However, we only consider the 2 last seconds of videos. So in our case, no recency effect can be assessed, but just whether any improvement in VQA performance can be seen by taking more or less video footage within the very final seconds of a video.

Minkowski summation, kth percentile and last seconds pooling have been used in the past literature, providing evidence that poling methods simulating a recency effect or emphasizing the effect of poor quality frames boost the performance of VQA algorithms (e.g. See Rimac-Drlje et al., 2009; Keimel and Diepold, 2010; Park et al., 2020. But see Seufert et al., 2013, who found no additional gain with respect to the arithmetic mean using their best performing pooling methods such as kth percentile).

2. Use of temporal information across frames by INRF

Let us recall for a moment the INRF model described in Eq. 2 from the manuscript and rename it here as Eq. 8:

$$\mathrm{INRF}(x) = m \star I(x) - \lambda \sum_i w_i \, \mathcal{S}\big(I(y_i) - g \star I(x)\big) \qquad (8)$$

where the kernels m, w, g are 2D Gaussians of standard deviations $\sigma_m, \sigma_w, \sigma_g$ respectively, λ is a scalar, $\mathcal{S}$ is a sigmoid that has the form of an **atan** function, and the symbol $\star$ denotes the 2D convolution.

The INRF equation just described operates on single images. Therefore, we may extend it to the temporal domain, representing it as $\mathrm{INRF}(x, f)$, where f is the video frame to be computed, to account for its operation over the different frames in a video.

However, the INRF model reported does not incorporate any temporal information present across video frames. What it does is perform its computations individually over each of them, without past frames having any influence on the video frame being processed at present.

To make INRF responsive to temporal information across frames, $\mathrm{INRF}(x, f)$ can be convolved with a temporal impulse response function as follows:

$$0_{tf}(x, f) = INRF_{tf}(x, f) = T \otimes INRF(x, f) \qquad (9)$$

where $\star$ denotes the convolution operation and T is a temporal impulse response function taken from the motion energy model by Adelson and Bergen (1985). This temporal response function has a positive and a negative lobe, in consonance with neurophysiological findings of temporal response properties of visual neurons (Fukushima et al., 2001; Theunissen et al., 2001; Reid and Shapley, 2002), and is defined as:

$$T(t) = (kt)^n e^{(-kt)} \left(\frac{1}{n!} - \frac{(kt)^2}{(n+2)^2} \right) \qquad (10)$$

where $k = 1$ and $n = 5$ and t stands for time. The total length in time for the temporal filter is chosen to be 100 ms (0.1 sec), in agreement with Reid and Shapley, 2002 (see their Figure 10A and Figure 11A, the responses of parvocellular and magnocellular neurons to luminance stimuli). Thus, if for instance, a 120 fps video were to be processed, the total length of the temporal filter would be 12 (i.e. $120\ Hz\ \times 0.1\ sec$).

The INRF-VQA value comparing a reference video V and its distorted version V_D is computed as:

$$\begin{aligned} INRF\text{-}VQA_{tf}(V, V_D) &= \mathrm{Tp}\left\{\sqrt{\mathrm{MSE}\left(0_{tf}, 0_{tf_D}\right)}\right\} \\ &= \mathrm{Tp}\left\{\sqrt{\mathrm{MSE}\left(INRF_{tf}(I^i, f), INRF_{tf}(I^i{}_D, f)\right)}\right\} \end{aligned} \tag{11}$$

where I^i and $I^i{}_D$ are, respectively, the i-th frame of V and V_D, and the $\mathrm{INRF_{tf}}$ has the same form as in Eq. 9. Be aware that the term $\sqrt{\mathrm{MSE}\left(INRF_{tf}(I^i, f), INRF_{tf}(I^i{}_D, f)\right)}$ outputs values in a frame-wise manner. So these also need to be temporally pooled, despite a temporal filtering stage having been implemented. In this regard, Tp stands for a temporal pooling strategy; any from the ones described in the section "1. Temporal pooling methods in INRF-VQA metrics" from this Supplementary Material can be used.

With temporal processing incorporated into INRF-VQA, and having named the VQA metric that does consider temporal information across frames $\mathrm{INRF\text{-}VQA_{tf}}$, we now refer to the VQA metric proposed in the manuscript, which does not consider temporal information, as $\mathrm{INRF\text{-}VQA_{no.tf}}$.

3. Supplementary Material Results

We use our INRF-VQA metrics, both INRF-VQAtf and INRF-VQAno.tf, over videos from the LIVE-YouTube-HFR dataset and correlate their objective video quality scores with subjective quality scores for those same videos.

Table S1 shows the resulting Spearman rank correlation coefficients (SRCC), Pearson correlation coefficients (PLCC) and Kendall correlation coefficients (KRCC) for 120 fps, 60 fps and 30 fps frame rates in the LIVE-YouTube-HFR dataset. Results are shown for the different temporal pooling methods described in this Supplementary Material. When it comes to Minkowski summation, last seconds and kth percentile, the table shows the maximum correlation value obtained out of the different parameter values used in those methods: for Minkowski summation, the exponent values tried are $\{0.1, 0.25, 0.5, 0.75, 1, 3, 5, 7, 10\}$; for $\mathrm{k^{th}}$ percentile, the percentiles used are $\{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 99\}$, and last seconds is used over the durations (in seconds) $\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$ (recall that our INRF-VQA metrics are only run over the last 2 seconds of videos). To finish with, INRF-VQAtf is only used over 120 fps videos. Frame dropping of reference videos is used on 60 fps and 30 fps in order to mach reference and distorted videos in their number of frames.

Table S1. Correlation results (SRCC, PLCC and KRCC) of video quality scores provided by INRF-VQA models with the corresponding subjective measurements of video quality are shown for different frame rates (120 fps, 60 fps and 30 fps). In the case of 120 fps, results are shown for INRF-$VQA_{no.tf}$ and INRF-VQA_{tf} models. For 60 fps and 30 fps, results are shown only for the INRF-$VQA_{no.tf}$ model. Results for different temporal pooling methods are displayed: arithmetic mean, median, maximum score, minimum score, Minkowski summation, last seconds and k^{th} percentile pooling. In the case of Minkowski summation, last seconds and k^{th} percentile pooling, the maximum correlation obtained is shown as well as the pooling parameter for which that correlation is achieved. For 60 fps and 30 fps, results are shown using frame dropping of reference videos.

	120 fps						**60 fps**			**30 fps**		
	INRF-VQA w. n. temporal filtering			**INRF-VQA w. temporal filtering**			**INRF-VQA w. n. temporal filtering**			**INRF-VQA w. n. temporal filtering**		
	SRCC	**PLCC**	**KRCC**	**SRCC**	**PLCC**	**KRCC**	**SRCC**	**PLCC**	**KRCC**	**SRCC**	**PLCC**	**KRCC**
Arithmetic mean	0.8452	0.8169	0.6468	0.8463	0.8162	0.6478	0.7515	0.8193	0.5639	0.6292	0.8715	0.4639
Median	0.8331	0.8083	0.6329	0.8349	0.8087	0.6359	0.7404	0.8108	0.5517	0.6191	0.8611	0.4529
Max value	0.8486	0.8209	0.6607	0.8475	0.8203	0.6587	0.7524	0.8126	0.5672	0.6279	0.8658	0.4581
Min value	0.8821	0.8659	0.6894	0.8821	0.8660	0.6894	0.7648	0.8628	0.5788	0.6514	0.9027	0.4891
Minkowski (max)	0.8457 (0.75)	0.8175 (0.1)	0.6478 (0.75)	0.8464 (0.5, 0.75)	0.8168 (0.1)	0.6488 (0.5, 0.75)	0.7527 (0.1, 0.25, 0.5)	0.8211 (0.1)	0.5652 (0.1, 0.25, 0.5)	0.6309 (0.25)	0.8741 (0.1)	0.4658 (0.25)
Last secs (max)	0.8455 (1.75 s)	0.8169 (2 s)	0.6478 (1.75 s)	0.8464 (1.75 s)	0.8162 (2 s)	0.6488 (1.75 s)	0.7547 (0.75 s)	0.8278 (0.25 s)	0.5691 (0.75 s)	0.6316 (0.5 s)	0.8795 (0.5 s)	0.4684 (0.5 s)
Percentile (max)	0.8492 (99%)	0.8204 (99%)	0.6607 (99%)	0.8478 (99%)	0.8198 (99%)	0.6577 (99%)	0.7538 (85%)	0.8182 (5%)	0.5672 (75%, 99%)	0.6290 (90%)	0.8699 (5%)	0.4607 (5%)

Figure S1 shows the results obtained when the temporal pooling methods used are Minkowski summation, last seconds and kth percentile pooling. Results are shown for the different parameter values used in the pooling methods. Only SRCC and PLCC correlations are displayed, as KRCC correlations show a similar tendency.

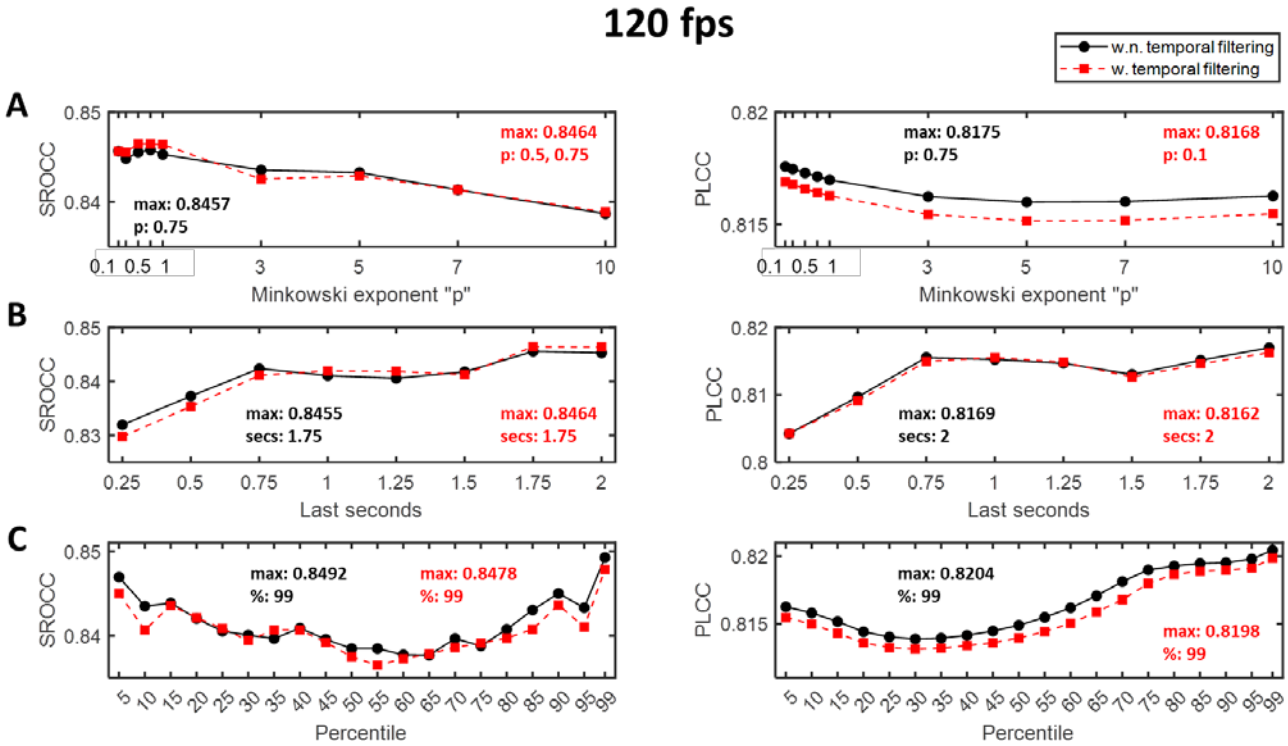

Figure S1**.** Correlations of video quality scores provided by INRF-VQA models with the corresponding subjective measurements of video quality. Correlation results are shown for: **(A)** different values of the *p* exponent from Minkowski summation {0.1, 0.25, 0.5, 0.75, 1, 3, 5, 7, 10} used for temporal pooling, **(B)** different durations of the last seconds of video taken for temporal pooling, **(C)** different percentiles in the kth percentile pooling method. SRCC and PLCC values are shown on the left and right panels respectively. Results obtained using an INRF-VQA model that considers temporal information across frames (INRF-VQA$_{tf}$) and one that does not (INRF-VQA$_{no.tf}$) are shown as black circles and red squares with dashed lines respectively. Maximum correlation values are displayed in the plots specifying the corresponding parameter values for which those maximal correlations are obtained (Black font: INRF-VQA$_{no.tf}$. Red font: INRF-VQA$_{tf}$). Be aware that correlation axes may be different in each panel.

4. Supplementary Material References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *JOSA A*, *2*(2), 284-299.

Batsi, S., & Kondi, L. P. (2020). Improved Temporal Pooling for Perceptual Video Quality Assessment Using VMAF. *Electronic Imaging*, *2020*(11), 68-1.

Fukushima, Y., Hara, K., & Kimura, M. (1986). A spatio-temporal model of ganglion cell receptive field in the cat retina. *Biological cybernetics, 54*(2), 91-98.

Kaplan, E., & Benardete, E. (2001). The dynamics of primate retinal ganglion cells. *Progress in brain research*, 134, 17-34.

Keimel, C., & Diepold, K. (2010). Improving the prediction accuracy of PSNR by simple temporal pooling. *In Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics-VPQM 2010*.

Park, J., Seshadrinathan, K., Lee, S., & Bovik, A. C. (2012). Video quality pooling adaptive to perceptual distortion severity. *IEEE Transactions on Image Processing, 22*(2), 610-620.

Reid, R. C., & Shapley, R. M. (2002). Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus. *Journal of Neuroscience, 22*(14), 6158-6175.

Rimac-Drlje, S., Vranjes, M., & Zagar, D. (2009, May). Influence of temporal pooling method on the objective video quality evaluation. *In 2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting* (pp. 1-5). IEEE.

Seufert, M., Slanina, M., Egger, S., & Kottkamp, M. (2013, July). "To pool or not to pool": A comparison of temporal pooling methods for HTTP adaptive video streaming. *In 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)* (pp. 52-57). IEEE.

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems, 12*(3), 289.