Capturing Video Frame Rate Variations via Entropic Differencing

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik

***Abstract*—High frame rate videos are increasingly getting popular in recent years, driven by the strong requirements of the entertainment and streaming industries to provide high quality of experiences to consumers. To achieve the best trade-offs between the bandwidth requirements and video quality in terms of frame rate adaptation, it is imperative to understand the effects of frame rate on video quality. In this direction, we devise a novel statistical entropic differencing method based on a Generalized Gaussian Distribution model expressed in the spatial and temporal band-pass domains, which measures the difference in quality between reference and distorted videos. The proposed design is highly generalizable and can be employed when the reference and distorted sequences have different frame rates. Our proposed model correlates very well with subjective scores in the recently proposed LIVE-YT-HFR database and achieves state of the art performance when compared with existing methodologies.**

***Index Terms*—High frame rate, video quality assessment, full reference, entropy, natural video statistics, generalized Gaussian distribution.**

I. Introduction

As CURRENT media technology continues to emphasize ever higher quality regimes and to involve more immersive and engaging experiences for consumers, the need to extend current video parameter spaces along spatial and temporal resolutions, screen sizes and dynamic ranges has become a topic of extreme importance, especially in the media and streaming industry. Existing and emerging standards have increasingly focused on improving spatial resolution (4K/8K) [1], High Dynamic Range (HDR) [2], [3], and multiview formats [4], [5]. However there has been much less emphasis placed on increasing frame rates, and for a long time the frame rates associated with television, cinema and other video streaming applications have changed little - rarely above 60 frames per second (fps).

Various factors have limited increased adoptions of High Frame Rate (HFR) videos. Switching to HFR requires employing complex capture and display technologies which were not commonly available. Another possible reason for the limited popularity of HFR relates to limited knowledge about the perceptual benefits of employing HFR, which partly arises due to insufficient availability of HFR content. Recently, HFR has gathered significant interest among the research community, along with publication of databases such as the Waterloo HFR [6], BVI-HFR [7] and LIVE-YT-HFR [8] datasets that exclusively target HFR contents.

Perceptual Video Quality Assessment (VQA) is an integral component in numerous video applications such as digital cinema, video streaming applications (such as YouTube, Netflix, Hulu etc.) and social media (Facebook, Instagram etc). VQA models can be broadly classified into three main categories [9]: Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) models. FR VQA models require entire pristine undistorted stimuli along with degraded versions [10]–[16], while RR models operate with limited reference information [17]–[21]. NR models operate without any knowledge of pristine stimuli [22]–[25]. This work addresses the problem of quality evaluation when pristine and distorted sequences can possibly have different frame rates, thus our primarily focus will be on FR and RR VQA methods.

There has been very limited work done on addressing VQA in the HFR domain. One of the first models was proposed by Nasiri *et al.* [26], where they measured the amount of aliasing occurring in the temporal frequency spectrum, employing that as a measure of quality. In [27] a motion smoothness measure was proposed for cross frame rate quality evaluation. Zhang *et al.* [28] proposed a wavelet domain based Frame Rate Quality Metric (FRQM), where the differences between the wavelet coefficients of reference and temporally upsampled distorted sequences were used to predict quality. FRQM has a restriction that it cannot be employed when the reference and distorted videos have same frame rate, thus limiting its generalizability.

In this letter, we propose a statistical VQA model that can capture distortions arising due to frame rate variations, and provide quality predictions that correlate well with human perception. This model is primarily motivated by temporal variations observed in the distributions of band-pass coefficients. We propose a novel entropic differencing method using Generalized Gaussian Distribution (GGD) model for both spatial and temporal band-pass responses, and show its effectiveness in capturing spatio-temporal artifacts. We evaluate our model on the LIVE-YT-HFR database and show that the predicted quality estimates have superior correlations against human judgments as compared to existing methods. Our proposed method is simplistic in nature, has very few hyperparameters to tune and does not require any computationally intensive training process.

The rest of the letter is organized as follows. In Section II we provide a detailed description of our proposed VQA model. In

Manuscript received July 31, 2020; revised September 17, 2020; accepted September 22, 2020. Date of publication October 5, 2020; date of current version October 19, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sumohana S. Channappayya. *(Corresponding author: Pavan C. Madhusudana.)*

Pavan C. Madhusudana and Alan C. Bovik are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: pavancm@utexas.edu; bovik@ece.utexas.edu).

Neil Birkbeck, Yilin Wang, and Balu Adsumilli are with Google Inc., Mountain View, CA 94043 USA (e-mail: birkbeck@google.com; yilin@google.com; badsumilli@google.com).

Digital Object Identifier 10.1109/LSP.2020.3028687

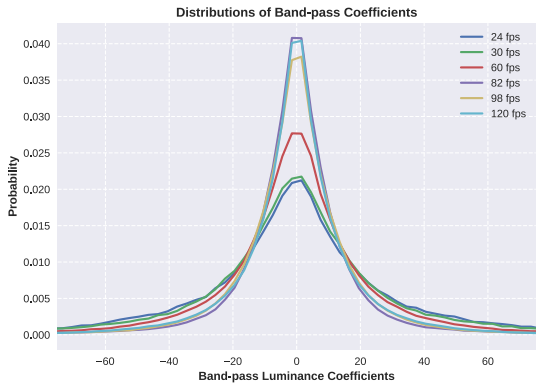

Fig. 1. Distributions of band-pass coefficients for six different frame rates.

Section III we report and analyze various experimental results, and provide some concluding remarks in Section IV.

II. Proposed Method

Consider a bank of K temporal band-pass filters denoted by b_k for $k \in \{1, \dots K\}$, the temporal band-pass response for a video $V(\mathbf{x}, t)$ ($\mathbf{x} = (x, y)$ represents spatial co-ordinates and t denotes temporal dimension) is given by

$$B_k(\mathbf{x}, t) = V(\mathbf{x}, t) * b_k(t) \quad \forall k \in \{1, \dots K\}, \tag{1}$$

where B_k denotes band-pass response of k^{th} filter. Note that these are 1D filters applied only along the temporal dimension. We empirically observe that the distributions of the coefficients of B_k vary as a function of frame rate. This is illustrated in Fig. 1, where distributions at different frame rates are shown for a 4-level Haar wavelet filter. From Fig. 1 it maybe observed that as frame rates increase, the distribution becomes more peakier as the correlation between the consecutive frames increase with frame rate. Since coefficients of B_k are band-pass in nature, they can be well modelled as following a Generalized Gaussian Distribution (GGD). GGD models have been widely used to model band-pass coefficients in many previous applications, such as image denoising [29], texture retrieval [30] etc. In this work we propose to employ entropic differences of band-pass GGD samples to quantify the deviations in distribution of band-pass coefficients.

A. GGD Based Statistical Model

Let the reference and distorted videos be denoted by R and D respectively, with R_t, D_t representing corresponding frames at time t. Note that R and D can have different frame rates though we require them to have same spatial resolution. Let the response of the k^{th} band pass filter b_k, $k \in \{1, 2, \dots K\}$ on reference and distorted videos be denoted by B^R_{kt} and B^D_{kt} respectively. We assume that every frame of B^R_{kt}, B^D_{kt} follows a GGD model with zero-mean. We divide each frame into P spatial blocks each of size $\sqrt{M} \times \sqrt{M}$. Let B^R_{kpt} and B^D_{kpt} denote vector of band pass coefficients in block p for subband k and frame t for reference and distorted respectively. We allow the band-pass coefficients to pass through a Gaussian channel to model perceptual imperfections such as neural noise [12], [20]. Let $\tilde{B}^R_{kpt}, \tilde{B}^D_{kpt}$ represent coefficients which undergo channel imperfections to obtain the observed responses B^R_{kpt}, B^D_{kpt} respectively. Also let $\tilde{B}^R_{kpt}, \tilde{B}^D_{kpt}$ both be modeled as following GGD. This model is expressed as:

$$B^R_{kpt} = \tilde{B}^R_{kpt} + W^R_{kpt} \quad B^D_{kpt} = \tilde{B}^D_{kpt} + W^D_{kpt} \tag{2}$$

where $\tilde{B}^R_{kpt}$ is independent of W^R_{kpt}, $\tilde{B}^D_{kpt}$ is independent of W^D_{kpt}, and where W^R_{kpt}, W^D_{kpt} are drawn from the Gaussian distribution $\mathcal{N}(0, \sigma^2_W \mathbf{I_M})$. It can be inferred from (2) that B^R_{kpt}, B^D_{kpt} need not necessarily be GGD, although it can be well approximated by a GGD [31] due to the independence assumption. As shown in the VIF [12] formulation, distortion results in a loss of "natural" image information as measured by suitably defined entropies. Variations over time of video frames from distortion can affect this visual information flow, and may depend on frame rate. For example, a lower frame rate may result in judder, which measurably affects the information flow, as measured by entropy under the statistical model of videos. The entropy of a GGD random variable $X \sim GGD(0, \alpha, \beta)$ has a closed form expression given by:

$$h(X) = \frac{1}{\beta} - \log\left(\frac{\beta}{2\alpha\Gamma(1/\beta)}\right) \tag{3}$$

where α and β are the scale and shape parameters of GGD respectively. Entropy computation requires the values of the GGD parameters of $\tilde{B}^R_{kpt}$ and $\tilde{B}^D_{kpt}$. However we only have access to B^R_{kpt} and B^D_{kpt}. In order to estimate these parameters we follow the kurtosis matching procedure detailed in [32] from which kurtosis values of $\tilde{B}^R_{kpt}$and $\tilde{B}^D_{kpt}$ can be obtained. The GGD parameters and kurtosis follow a bijective mapping [32] where the kurtosis of a GGD random variable is given by:

$$Kurtosis(X) = \frac{\Gamma(5/\beta)\Gamma(1/\beta)}{\Gamma(3/\beta)^2} \tag{4}$$

A simple grid search can be used to estimate the shape parameter β from obtained kurtosis value. The other parameter α can be obtained using the relation

$$\alpha = \sigma\sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}} \tag{5}$$

Plugging the parameters obtained from (4) and (5) in (3), the entropies $h(\tilde{B}^R_{kpt})$ and $h(\tilde{B}^D_{kpt})$ can be computed. In the next section we show how these entropies can be effectively used to assess the quality of videos.

B. Temporal Measure

We define entropy scaling factors given by:

$$\gamma^R_{kpt} = \log(1 + \sigma^2(\tilde{B}^R_{kpt})), \quad \gamma^D_{kpt} = \log(1 + \sigma^2(\tilde{B}^D_{kpt}))$$

These scaling factors are similar to the ones used in [19], [20]. Scaling factors lend a more local nature to our model and provide numerical stability on regions having low variance, where the entropy estimates are less stable. The entropies are modified by premultiplying with the scaling factors as shown in (6). Regions having low variances will have small scaling factors, reducing the impact of noise on the entropy values:

$$\epsilon^R_{kpt} = \gamma^R_{kpt} h(\tilde{B}^R_{kpt}), \quad \epsilon^D_{kpt} = \gamma^D_{kpt} h(\tilde{B}^D_{kpt}). \tag{6}$$

There exists a frame rate bias associated with the entropy values where different frame rates have entropies at different *scales*. High frame rate sequences such as 120 fps have much lower

TABLE I
PERFORMANCE COMPARISON OF FR-VQA ALGORITHMS ON THE HFR DATABASE. IN EACH COLUMN THE FIRST AND SECOND BEST VALUES ARE BOLDFACED AND UNDERLINED, RESPECTIVELY

	SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓
PSNR	0.6950	0.5071	0.6685	9.023
SSIM [10]	0.4494	0.3102	0.4526	10.819
MS-SSIM [11]	0.4898	0.3407	0.4673	10.726
FSIM [13]	0.4469	0.3151	0.4435	10.874
ST-RRED [20]	0.5531	0.3800	0.5107	10.431
SpEED [21]	0.4861	0.3409	0.4449	10.866
FRQM [28]	0.4216	0.2956	0.452	10.804
VMAF [34]	<u>0.7303</u>	<u>0.5358</u>	<u>0.7071</u>	<u>8.587</u>
deepVQA [35]	0.3463	0.2371	0.3329	11.441
GSTI (Ours)	**0.7909**	**0.5979**	**0.7910**	**7.422**

entropy values when compared to lower frame rates such as 24 fps, 30 fps etc. Thus simple entropy subtraction measures the difference between the frame rates of R and D. Though this is desirable, this can be inefficient when comparing videos which only differ by compression artifacts. To remove this bias, we employ an additional video sequence termed Pseudo Reference (PR) signal, which is obtained by temporally downsampling the reference to match the frame rate of the distorted video. In our implementation we use frame dropping to conduct temporal downsampling using the FFmpeg [33] tool. In the case when the distorted sequence has the same frame rate as the reference, PR will be the same as R. Similar to ϵ^R_{kpt} and ϵ^D_{kpt}, we calculate ϵ^{PR}_{kpt}. We define the Generalized Temporal Index (GTI) as:

$$GTI_{kt} = \frac{1}{P}\sum_{p=1}^{P}\left|\left(1+|\epsilon^D_{kpt}-\epsilon^{PR}_{kpt}|\right)\frac{\epsilon^R_{kpt}+1}{\epsilon^{PR}_{kpt}+1}-1\right|. \quad (7)$$

(7) can be interpreted by decomposing into two factors: absolute difference term and ratio term. Absolute difference term removes frame rate bias and captures the quality changes as if R and D were at the same frame rate. The ratio term weights these factors depending on the reference and distorted frame rate. In the case of reference and distorted videos having same frame rate, the ratio term will be 1, thus making (7) depend only on absolute difference. The unit terms within the absolute values ensure that GTI does not become zero when $D = PR \neq R$, which happens when distorted sequence is temporally subsampled version of the reference. Note that $GTI = 0$ only when $D = PR = R$. The unit terms in the ratio avoid indeterminate values in regions having small entropy values.

C. Spatial Measure

Although GTI does capture spatial information due to its spatial block based nature, it is primarily influenced by the temporal filtering. To extract information about spatial artifacts, we employ spatial band-pass filters applied to every frame of the video. For this purpose we employ a local Mean Subtracted (MS) filtering similar to [21]. Let $R^{MS}_t = R_t - \mu^R_t$ and $D^{MS}_t = D_t - \mu^D_t$ be the reference and distorted MS coefficients where local mean is calculated as

$$\mu^R_t(i,j) = \sum_{g=-G}^{G}\sum_{h=-H}^{H}\omega_{g,h}R_t(i+g,j+h),$$

$$\mu^D_t(i,j) = \sum_{g=-G}^{G}\sum_{h=-H}^{H}\omega_{g,h}D_t(i+g,j+h)$$

where $\omega = \omega_{g,h}|g=-G,\dots G, h=-H,\dots H$ is a 2D circularly symmetric Gaussian weighting function sampled out to 3 standard deviations. In our implementation we use $G = H = 7$. The MS coefficients R^{MS}_t, D^{MS}_t are modeled as following a GGD model. Similar to the temporal measure, we divide each frame into P nonoverlapping blocks and calculate entropies $h(\tilde{R}^{MS}_t)$ and $h(\tilde{D}^{MS}_t)$ as detailed in Subsection II-A by replacing temporal band-pass responses with corresponding MS coefficients. Similarly we define scaling factors and modified entropies:

$$\eta^R_{pt} = \log(1+\sigma^2(\tilde{R}^{MS}_{pt})), \quad \eta^D_{pt} = \log(1+\sigma^2(\tilde{D}^{MS}_{pt}))$$
$$\theta^R_{pt} = \eta^R_{pt}h(\tilde{R}^{MS}_{pt}), \qquad \theta^D_{pt} = \eta^D_{pt}h(\tilde{D}^{MS}_{pt}).$$

Since spatial entropies are computed using only the information from a single frame, the values are frame rate agnostic. Thus there does not arise any scale variations due to frame rate, as seen in the temporal case. The Generalized Spatial Index (GSI) is then defined as:

$$GSI_t = \frac{1}{P}\sum_{p=1}^{P}|\theta^D_{pt}-\theta^R_{pt}|. \quad (8)$$

D. Spatio-Temporal Measure

GSI and GTI operate individually on data obtained by separate processing of spatial and temporal frequency responses. Interestingly, while GSI is obtained in a purely spatial manner, GTI has both spatial and temporal information embedded in it (as entropies are obtained in a spatial blockwise manner). Thus temporal artifacts such as judder etc. only influence GTI, while spatial artifacts affect both GTI and GSI. A combined Generalized Spatio-Temporal Index (GSTI) is defined as:

$$GSTI_{kt} = GTI_{kt}GSI_t. \quad (9)$$

The quality score obtained from (9) provides scores at frame level. To obtain a video level quality score we average pool (tacitly assuming frames are temporally consistent, i.e., do not contain scene cuts, which are easily detected) the frame scores:

$$GSTI_k = \frac{1}{T}\sum_{t=1}^{T}GSTI_{kt}. \quad (10)$$

Implementation Details: For simplicity we implemented our method only in the luminance domain. We use a 3-level Haar wavelet filter as the temporal band-pass filter b_k with $k \in \{1,\dots 7\}$ (we ignore the low pass response), where a higher k value denotes a larger center frequency. We used wavelet packet (constant linear bandwidth) (WP) filter bank [36] as we found it to be more effective than using constant octave bandwidth filters. For entropy calculation we choose spatial blocks of size 5×5 (i.e., $\sqrt{M} = 5$). We choose neural noise variance $\sigma^2_W = 0.1$ defined in (2). Note that similar values were employed in [12]

TABLE II
PERFORMANCE COMPARISON OF VARIOUS FR METHODS FOR INDIVIDUAL FRAME RATES IN THE LIVE-YT-HFR DATABASE. IN EACH COLUMN FIRST AND SECOND BEST VALUES ARE BOLDFACED AND UNDERLINED, RESPECTIVELY

	24 fps		30 fps		60 fps		82 fps		98 fps		120 fps		Overall	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
PSNR	0.4101	0.3647	0.4414	0.4179	0.6202	0.5719	0.6878	0.6431	0.7171	0.6489	0.6019	0.5937	0.6950	0.6685
SSIM [10]	0.1277	0.0949	0.1108	0.0816	0.2123	0.1845	0.2079	0.2430	0.3876	0.3964	0.7485	0.6726	0.4494	0.4526
MS-SSIM [11]	0.2221	0.1500	0.1929	0.1112	0.2516	0.1900	0.2906	0.2549	0.4237	0.4007	0.6165	0.5843	0.4898	0.4673
FSIM [13]	0.3670	0.3038	0.3208	0.2638	0.2472	0.2615	0.3225	0.3055	0.3861	0.2646	0.3056	0.1178	0.4469	0.4435
ST-RRED [20]	0.1541	0.0369	0.1188	0.0307	0.5062	0.4457	0.3394	0.3271	0.4962	0.4556	0.6745	0.5906	0.5531	0.5107
SpEED [21]	0.2591	0.1237	0.2278	0.0896	0.1824	0.1110	0.2955	0.2425	0.4118	0.3295	0.6827	0.6097	0.4861	0.4449
FRQM [28]	0.1556	0.2089	0.0983	0.0854	0.0947	0.0309	0.0137	0.0035	0.0317	0.0100	-	-	0.4216	0.4520
VMAF [34]	0.1743	0.2669	0.2855	0.3740	0.5408	0.6015	0.6820	0.7390	**0.8214**	**0.8128**	**0.7943**	**0.7844**	0.7303	0.7071
deepVQA [35]	0.1144	0.0495	0.1353	0.1059	0.2527	0.1652	0.1803	0.1515	0.2816	0.2654	0.6865	0.6209	0.3463	0.3329
GSTI (Ours)	**0.4538**	**0.5935**	**0.4758**	**0.6689**	**0.6552**	**0.7566**	**0.7633**	**0.8183**	0.7844	0.7775	0.7390	0.7003	**0.7909**	**0.7910**

and [19]. We observed that our algorithm is most effective when spatial resolution is downsampled 16 times along both dimensions. Similar observations were made in [20] and [21] and is attributed to the motion downshifting phenomenon where, in presence of motion, human vision tends to be more sensitive to coarser scales than finer ones. Since reference and distorted sequences can have different frame rates, the reference entropy terms ϵ^R_{kpt}, θ^R_{kt} will have a different number of frames when compared to their counterpart distorted entropy terms ϵ^D_{kpt}, θ^D_{pt}. Thus we temporally average reference entropy terms as:

$$\epsilon^R_{kpt} \leftarrow \frac{1}{F}\sum_{n=1}^{F}\epsilon^R_{kpt'} \quad \text{where} \begin{cases} F &= \frac{FPS_{\text{ref}}}{FPS_{\text{dist}}}, \\ t' &= (t-1)F + n \end{cases}$$

$$\theta^R_{pt} \leftarrow \frac{1}{F}\sum_{n=1}^{F}\theta^R_{pt'}$$

III. EXPERIMENTS

Experimental Settings: We selected 4 FR-IQA methods: PSNR, SSIM [10], MS-SSIM [11] and FSIM [13] for comparison. Since these are image indices, they are computed on every frame and averaged across all frames to obtain the video scores. In addition to the above IQA indices, we also include 5 FR-VQA indices: ST-RRED [20], SpEED [21], FRQM[28], VMAF[1] [34] and deepVQA [35]. For deepVQA, we use only stage-1 of the pretrained model (trained on the LIVE-VQA [37] database) obtained from the code released by the authors. Since the above methods require same frame rates for reference and distorted videos, for cases with differing frame rates, the distorted video was temporally upsampled by frame duplication to match the reference frame rate. Although we can downsample the reference as well, we avoided this method since it can potentially introduce artifacts (e.g., judder) in the reference video which is not desirable. All the above VQA models were evaluated at their original spatial resolution. Spearman's rank order correlation coefficient (SROCC), Kendall's rank order correlation coefficient (KROCC), Pearson's linear correlation coefficient (PLCC) and root mean squared error (RMSE) were the main performance criteria employed to evaluate the VQA methodologies. Before computing PLCC and RMSE, the predicted scores were passed through a four-parameter logistic non-linearity, as described in [38].

[1]We use the pretrained VMAF model available at https://github.com/Netflix/vmaf

A. Correlation Against Human Judgments

The correlations between objective scores predicted by various FR models against the human judgments in the LIVE-YT-HFR database are compared in Table I. Our proposed method outperformed all the existing models across every evaluation criteria, as illustrated in Table I. The reported results for GSTI in Table I correspond to the first subband (i.e., b_1) of the band-pass filter, which was empirically observed to achieve highest performance when compared to other subbands.

B. Performance Analysis With Individual Frame Rates

In this experiment we subdivided the LIVE-YT-HFR database into sets which contain videos having the same frame rate, and individually analyzed the performance on them. The performance comparison is shown in Table II. To avoid clutter we only include SROCC and PLCC for evaluation. At high frame rates, there are naturally reduced temporal distortions, hence distortions are primarily from compression, which VMAF is (Pareto) optimized to handle. We also observed an interesting anomaly where PSNR achieved higher performance at lower frame rates when compared to other prior VQA models, which is surprising, since PSNR correlates poorly against human quality perception [39]. It is possible that frame-based models like SSIM, which accurately predict spatial distortions, have a "spatial bias" on this database. PSNR, which is merely a space-time difference signal will not have such a bias. For FRQM, correlation values are not reported for 120 fps, as it requires the compared videos to have different frame rates. It should be noted that a factor in the performance of FRQM (Table II) could be that it was designed on frame averaging, rather than frame dropping.

IV. CONCLUSION AND FUTURE WORK

We presented a simple, highly generalizable video quality evaluation method that can be employed when reference and distorted videos having different frame rates, and gauged its performance on the new LIVE-YT-HFR database. We performed a holistic evaluation of our method in terms of correlation against human perception and established that our method is superior and more robust than existing algorithms.

For band-pass filtering, a simple Haar filter was used, which can potentially limit performance. As part of future work we plan to explore other band-pass filters with superior frequency responses. Another avenue we wish to explore is to incorporate GSTI into a data driven quality model such as VMAF [34], to further enhance performance.

REFERENCES

[1] C. Ge, N. Wang, G. Foster, and M. Wilson, "Toward QoE-assured 4k video-on-demand delivery through mobile edge virtualization with adaptive prefetching," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2222–2237, Oct. 2017.

[2] Z. Mai, H. Mansour, R. Mantiuk, P. Nasiopoulos, R. Ward, and W. Heidrich, "Optimizing a tone curve for backward-compatible high dynamic range image and video compression," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1558–1571, Jun. 2011.

[3] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, Jun. 2017.

[4] A. Smolic *et al.*, "Coding algorithms for 3DTV—A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1606–1621, Nov. 2007.

[5] V. De Silva, H. K. Arachchi, E. Ekmekcioglu, and A. Kondoz, "Toward an impairment metric for stereoscopic video: A full-reference video quality metric to assess compressed stereoscopic video," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3392–3404, Sep. 2013.

[6] R. M. Nasiri, J. Wang, A. Rehman, S. Wang, and Z. Wang, "Perceptual quality assessment of high frame rate video," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2015, pp. 1–6.

[7] A. Mackin, F. Zhang, and D. R. Bull, "A study of high frame rate video formats," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1499–1512, Jun. 2019.

[8] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," 2020, *arXiv:2007.11634*.

[9] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[11] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[12] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[13] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[14] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[15] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2505–2508.

[16] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.

[17] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proc. Human Vis. Electron. Imag. X*, 2005, pp. 149–159.

[18] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[19] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, Feb. 2012.

[20] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[21] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[23] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[24] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[25] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, Jul. 2016.

[26] R. M. Nasiri and Z. Wang, "Perceptual aliasing factors and the impact of frame rate on video quality," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3475–3479.

[27] R. M. Nasiri, Z. Duanmu, and Z. Wang, "Temporal motion smoothness and the impact of frame rate variation on video quality," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 1418–1422.

[28] F. Zhang, A. Mackin, and D. R. Bull, "A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 300–304.

[29] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[30] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback–Leibler distance," *IEEE Trans. Image Process.*, vol. 11, no. 2, pp. 146–158, Feb. 2002.

[31] Q. Zhao, H.-W. Li, and Y.-T. Shen, "On the sum of generalized Gaussian random signals," in *Proc. IEEE Int. Conf. Signal Process.*, 2004, pp. 50–53.

[32] H. Soury and M.-S. Alouini, "New results on the sum of two generalized Gaussian random variables," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2015, pp. 1017–1021.

[33] FFmpeg, "Encoding for streaming sites." Accessed: Nov. 1, 2019. [Online]. Available: https://trac.ffmpeg.org/wiki

[34] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," 2016. [Online] Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[35] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 219–234.

[36] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 713–718, Mar. 1992.

[37] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[38] VQEG, "Final report from the Video Quality Experts Group on the validation of objective quality metrics for video quality assessment," 2000.

[39] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.