A COMPARATIVE EVALUATION OF TEMPORAL POOLING METHODS FOR BLIND VIDEO QUALITY ASSESSMENT

Zhengzhong Tu[1⋆], *Chia-Ju Chen*[1⋆], *Li-Heng Chen*[1], *Neil Birkbeck*[2], *Balu Adsumilli*[2], *and Alan C. Bovik*[1]

[1]The University of Texas at Austin, [2]YouTube Media Algorithms Team, Google Inc.

ABSTRACT

Many objective video quality assessment (VQA) algorithms include a key step of temporal pooling of frame-level quality scores. However, less attention has been paid to studying the relative efficiencies of different pooling methods on no-reference (blind) VQA. Here we conduct a large-scale comparative evaluation to assess the capabilities and limitations of multiple temporal pooling strategies on blind VQA of user-generated videos. The study yields insights and general guidance regarding the application and selection of temporal pooling models. In addition, we also propose an ensemble pooling model built on top of high-performing temporal pooling models. Our experimental results demonstrate the relative efficacies of the evaluated temporal pooling models, using several popular VQA algorithms evaluated on two recent large-scale natural video quality databases. Conclusively, we also provide an empirical recipe for applying temporal pooling of frame-based quality predictions.

Index Terms— Video quality assessment, temporal pooling, memory effect, visual attention, temporal visual masking

1. INTRODUCTION

Video quality assessment (VQA) models have been widely studied [1] as an increasingly important toolset used by the streaming and social media industries. While full-reference (FR) VQA research is gradually maturing and several algorithms [2, 3] are quite widely deployed, recent attention has shifted more towards creating better no-reference (NR) VQA models that can be used to predict and monitor the quality of authentically distorted user-generated content (UGC) videos. UGC videos, which are typically created by amateur videographers, often suffer from unsatisfactory perceptual quality, arising from imperfect capture devices, uncertain shooting skills, a variety of possible content processes, as well as compression and streaming distortions. In this regard, predicting UGC video quality is much more challenging than assessing the quality of synthetically distorted videos in traditional video databases. UGC distortions are more diverse, complicated, commingled, and no "pristine" reference is available.

Many researchers have proposed possible solutions to the blind VQA (BVQA) problem [4–10], among which a simple but reasonably effective strategy is to compute frame-level quality scores, e.g., as generated by image quality assessment (IQA) models, then to express the evolution or relative importance over time by applying temporal pooling on the frame-level quality scores. Simple temporal average pooling is a widely used scheme to augment BVQA models [5, 7, 10]. Other kinds of pooling that are used include harmonic mean [11], Minkowski mean [12, 13], percentile pooling [14, 15], and adaptively weighted sums [16, 17]. More sophisticated pooling strategies have considered memory effects, such as primacy, recency [12, 13, 18], and hysteresis [6, 9, 19, 20]. The general applicability of these pooling models, however, has not so far been deeply validated in the context of BVQA models for real-world UGC videos, though a few more directed studies have been conducted [12, 13, 21]. To date, no comprehensive studies have been conducted to establish the added values of the spectrum of available VQA pooling schemes.

Here we seek to help fill this gap by conducting a systematic evaluation of popular temporal pooling algorithms, as applied to leading NR IQA models on recently developed large scale UGC video quality databases. We assessed the benefits, generalizability, and stability of these pooling mechanisms. Our aim is to identify statistically verifiable pooling approaches that can be applied on top of future state-of-the-art IQA models to further produce consistently better predictions of video quality. We also propose an ensemble approach, wherein multiple pooling models are aggregated to deliver better retrospective quality prediction. Our experimental results demonstrate that the proposed ensemble pooling method reveals robustness among the top-performing models.

The rest of this paper is structured as follows. Section 2 summarizes previous related literature, while Section 3 describes details of the evaluated and proposed pooling algorithms. Experimental results and analysis are presented in Section 4, and finally, we conclude the paper in Section 5.

2. RELATED WORK

A variety of methods for spatial pooling of "quality-aware" features have been proposed and studied in [14, 22, 23], yet less effort has been applied to the study on temporal pooling

[⋆] Equal contribution

978-1-7281-6395-6/20/$31.00 ©2020 IEEE

methods for BVQA. The most related works to that reported here are the comparative evaluations of temporal pooling on short video clips [12, 21], and on longer adaptive streaming videos [13]. They have collectively included various pooling methods combined with several objective frame-level quality predictors, evaluated on different subjective databases. Among the studied temporal pooling methods are: simple averaging, percentile pooling [14], Minkowski pooling [12], harmonic mean pooling [11], and the more complex VQPooling scheme [16], which adaptively emphasizes the worst scores along the time dimension, wherein frame-level scores are clustered into two groups (low quality and high quality), then combined into a single score by upweighting low-quality scores. Methods like percentile and VQPooling are predicated by the accepted notion that quality judgments are heavily influenced by the worst parts of a video.

Another cognitive aspect relevant to temporal visual pooling is the serial-position effect (or memory effect) hypothesis [24]. Primacy and recency are two common effects that have been investigated in numerous video quality of experience (QoE) studies [18, 25, 26], but are less studied in regard to their influence on the blind quality prediction of UGC video clips. Another popular temporal memory modeling approach is hysteresis pooling [19], which has been justified in several video quality modeling papers [6, 9, 20]. The hysteresis model assumes that while subjective judgments drop sharply with event of poor video quality, they only recover slowly with subsequent improved video quality.

3. TEMPORAL QUALITY POOLING METHODS

We propose a comprehensive evaluation framework to study the influence of temporal pooling algorithms on the performances of objective video quality models. Suppose a video has N frames $\{F_1, F_2, ..., F_N\}$ processed by any NR IQA models that produces frame-level (time-varying) quality predictions $\{q_1, q_2, ...q_N\}$. The per-frame quality scores are temporally combined by a temporal pooling function $\mathcal{F}(\cdot)$ to obtain a final quality prediction: $Q_{\text{FINAL}} = \mathcal{F}(q_1, q_2, ..., q_N)$.

3.1. Frame Quality Prediction

Frame-level quality scores can be predicted by any NR IQA, such as BRISQUE [4], NIQE [27], FRIQUEE [8] or even models implemented as deep learning networks [28].

3.2. Temporal Pooling Models

Once frame-level quality scores $\{q_1, q_2, ...q_N\}$ are obtained, a variety of ways have been proposed to summarize the time-varying quality scores into a single overall video quality judgment. A variety of human factors have been explored in this context, including visual perception [29, 30], memory effects [18, 19, 25], and video content [9, 25, 31]. Here we model and study a collection of factors that express aspects of temporal quality perception, as candidates for deriving final quality predictions on UGC videos. Specifically, we study the following listed in approximate order of increasing complexity and abstraction:

Arithmetic Mean: The sample mean of frame-level scores is the most widely used method:

$$Q = \frac{1}{N}\sum_{n=1}^{N} q_n. \tag{1}$$

Harmonic Mean: The harmonic mean has been observed to emphasize the impact of low-quality frames [11]:

$$Q = \left(\frac{1}{N}\sum_{n=1}^{N} q_n^{-1}\right)^{-1}. \tag{2}$$

Geometric Mean: The third Pythagorean mean (geometric) expresses the central tendency of the quality scores by the product of their values:

$$Q = \left(\prod_{n=1}^{N} q_n\right)^{1/N}. \tag{3}$$

Minkowski Mean: The L_p Minkowski summation [12, 13] of time-varying quality is defined as:

$$Q = \left(\frac{1}{N}\sum_{n=1}^{N} q_n^p\right)^{1/p}. \tag{4}$$

Percentile: The idea of percentile pooling is based on observed phenomenon that perceptual quality is heavily affected by the "worst" parts of the content. Many prior works have studied and justified (or challenged) percentile pooling [12–15, 18]. The k-th percentile pooling is expressed:

$$Q = \frac{1}{|P_{\downarrow k\%}|}\sum_{n\in P_{\downarrow k\%}} q_n, \tag{5}$$

where $P_{\downarrow k\%}$ denotes the set of lowest $k\%$ scores.

VQPooling: VQPooling is an adaptive spatial and temporal pooling strategy proposed in [16]. Here we only study the temporal pooling part, wherein the quality scores of all frames are classified into two groups composed of higher and lower quality, using k-means clustering. The two groups, dubbed G_L and G_H, are then combined to obtain an overall quality prediction on the entire video sequence:

$$Q = \frac{\sum_{n\in G_L} q_n + w \cdot \sum_{n\in G_H} q_n}{|G_L| + w \cdot |G_H|}, \tag{6}$$

where $|G_L|$ and $|G_H|$ denote the cardinality of G_L and G_H, while the weight w is defined as the ratio between the scores in G_L and G_H:

$$w = \left(1 - \frac{M_L}{M_H}\right)^2, \tag{7}$$

where M_L and M_H are the average value of the quality scores in set G_L and G_H, respectively.
Temporal Variation: The approach of [32] considers the temporal changes of spatial distortions and proposes short-term and long-term spatiotemporal pooling mechanisms to account for quality changes. Here we only utilize the temporal variation terms in our study:

$$Q = \frac{1}{|P_{\uparrow k\%}|} \sum_{n \in P_{\uparrow k\%}} |q_n - q_{n-1}|, \tag{8}$$

where $|q_n - q_{n-1}|$ is the absolute quality difference at time n, and $P_{\uparrow k\%}$ is the set of largest $k\%$ absolute quality differences.
Primacy Effect: The primacy effect describes the tendency of human viewers to recall the earliest portion of a video when providing overall evaluations [24]. One way of capturing primacy is as an exponentially decreasing weighted sum. Define

$$Q = \sum_{n=1}^{N} w_n q_n, \tag{9}$$

where

$$w_n = \frac{\exp(-\alpha_p n)}{\sum_{k=0}^{L} \exp(-\alpha_p k)}, \; 0 \le n \le L, \tag{10}$$

where α_p is used to tune the intensity of primacy effect.
Recency Effect: The recency effect is another well-established behavioral and memory effect, whereby, in this context, video quality is very strongly influenced by a viewer's most recently percieved visual impression [24]. The recency effect can also be characterized as an exponential weighted sum (Eq. (9)), but with a different weighting:

$$w_n = \frac{\exp(-\alpha_r(L-n))}{\sum_{k=0}^{L} \exp(-\alpha_r(L-k))}, \; 0 \le n \le L, \tag{11}$$

where α_r tunes the relative intensity of the recency effect.
Temporal Hysteresis: This approach was inspired by the hysteresis effect observed in human judgments of time-varying video quality [19], which is closely related to, but not the same as the recency effect. The hysteresis measurement can be formulated as follows. Let $q_n, \; n = 1, 2, ...N$ be the time-varying frame quality scores. The memory of past quality l_n at the n-th frame is expressed as the minimum quality scores over the previous frames:

$$l_n = \begin{cases} q_n, & n = 1 \\ \min\limits_{k \in \mathcal{K}_{prev}} \{q_k\}, & n > 1, \end{cases} \tag{12}$$

where $\mathcal{K}_{prev} = \{\max\{1, n-\tau\}, ..., n-2, n-1\}$ indexes the previous τ frames. The current video quality m_n is expressed as a weighted sum of ordered [33] frame-level qualities:

$$\boldsymbol{v} = sort(\{q_k\}), \; k \in \mathcal{K}_{next}, \tag{13}$$

$$m_n = \sum_{j=1}^{J} v_j w_j, \; J = |\mathcal{K}_{next}|, \tag{14}$$

where $\mathcal{K}_{next} = \{n, n+1, ..., \min\{n+\tau, N\}\}$ indexes the next τ frames and $\{w_j\}$ is the descending half of a Gaussian weighting function. Linearly combining the memory and the current quality components in (12) and (14) yields time-varying scores that capture the hysteresis effect. The pooled video quality Q is computed as the global temporal average of the time-varying hysteresis-transformed predictions:

$$q'_n = \alpha m_n + (1-\alpha) l_n, \tag{15}$$

$$Q = \frac{1}{N} \sum_{n=1}^{N} q'_n, \tag{16}$$

where α adjusts the contributions of these two elements.

3.3. Ensemble Temporal Pooling

We have just described a diverse set of temporal pooling mechanisms, each either heuristically, statistically defined, or motivated by psychovisual reasoning. As might be expected, and as we shall show, the performances of these methods differ, and also vary on different datasets. Given that these methods likely capture different aspects of perceptual pooling, ensemble learning is a direct way to combine them towards creating a more reliable and generic quality predictor. We denote this ensemble-based temporal pooling as **EPooling**. Similar concepts of model fusion/ensemble have been successfully utilized on the IQA/VQA problems [3, 36, 37].

Suppose the quality scores delivered by a set of pooling methods are denoted $Q_i, \; i = 1, ..., I$, where I is the number of input model predictions. Then train an ensemble regressor to fuse the multiple predicted labels into a single final score:

$$Q_{\text{EPooling}} = \mathcal{F}(\boldsymbol{Q}), \; \boldsymbol{Q} = \{Q_i\}, \; i = 1, 2, ..., I, \tag{17}$$

where $\boldsymbol{Q}$ is the quality vector stacked from multiple singly pooled scores, and $\mathcal{F}$ is the learned regression function that maps the proxy quality vector to a final quality prediction Q_{EPooling}. Here we empirically chose Mean, VQPooling, and Hysteresis, as the three input prediction models after coarse preliminary feature analysis. Further improvements may be achieved by applying finer feature selection techniques.

4. EXPERIMENTS

4.1. Experimental Setup

We selected five popular NR IQA models: NIQE [27], BRISQUE [4], GM-LOG [38], HIGRADE [39], and CORNIA [40], as frame-level quality predictors, and evaluated the temporal pooling methods on two recent large scale UGC VQA databases: KoNViD-1k [34] and LIVE-VQC [35]. KoNViD-1k consists of 1,200 8-second 540p public-domain

Table 1: Performance comparison of temporal pooling methods as evaluated on KoNViD-1k [34] and LIVE-VQC [35]. Each cell shows the median evaluation results formatted as SRCC/PLCC. The three best results along each column are **boldfaced**.

Database	KoNViD-1k					LIVE-VQC				
Pool/Model	NIQE	BRISQUE	GMLOG	HIGRADE	CORNIA	NIQE	BRISQUE	GMLOG	HIGRADE	CORNIA
Mean	0.552/**0.560**	0.673/0.676	0.662/0.671	0.690/0.696	**0.749/0.764**	0.600/0.631	0.597/0.632	0.575/0.618	0.532/0.570	0.694/0.743
Median	0.543/0.554	0.667/0.670	0.657/0.666	0.680/0.689	**0.750/0.760**	0.584/0.618	0.577/0.619	0.558/0.602	0.521/0.559	0.687/0.744
Harmonic	0.550/**0.560**	**0.674**/0.676	0.667/0.674	0.693/0.699	0.696/0.696	0.607/0.637	0.605/0.636	0.585/0.620	0.537/0.575	**0.709**/0.737
Geometric	0.551/**0.560**	**0.676/0.679**	0.666/0.673	0.692/0.698	0.747/**0.760**	0.604/0.634	0.600/0.631	0.578/0.617	0.537/0.573	0.698/**0.746**
Minkowski	0.552/0.559	0.672/0.676	0.661/0.670	0.689/0.695	0.736/0.746	0.597/0.628	0.596/0.630	0.574/0.615	0.538/0.569	0.688/0.739
Percentile	0.545/0.547	0.655/0.647	**0.674/0.678**	0.685/0.687	0.696/0.700	**0.630**/0.634	**0.629/0.647**	**0.606/0.627**	**0.586/0.610**	**0.712**/0.744
VQPooling	0.549/0.554	0.670/0.665	**0.672**/0.674	**0.698/0.701**	0.743/0.758	**0.628/0.644**	**0.617/0.658**	**0.605/0.633**	0.563/0.597	0.700/**0.753**
Variation	0.347/0.328	0.348/0.338	0.509/0.511	0.434/0.444	0.240/0.303	0.507/0.476	0.470/0.463	0.495/0.488	0.474/0.482	0.567/0.609
Primacy	0.541/0.552	0.668/0.671	0.647/0.653	0.684/0.690	0.726/0.741	0.601/0.631	0.573/0.627	0.575/0.613	0.535/0.561	0.684/0.737
Recency	**0.553**/0.558	0.670/0.667	0.660/0.667	0.690/0.694	0.745/0.754	0.584/0.615	0.586/0.626	0.561/0.599	0.518/0.555	0.670/0.729
Hysteresis	**0.563/0.569**	**0.684/0.681**	**0.681/0.684**	**0.703/0.707**	0.732/0.735	0.621/**0.638**	**0.621/0.650**	**0.600/0.629**	**0.570/0.595**	**0.711/0.756**
EPooling	**0.572/0.579**	0.670/**0.679**	0.670/**0.676**	**0.698/0.704**	**0.749/0.762**	**0.623/0.645**	**0.617**/0.646	**0.605**/0.623	**0.582/0.601**	0.705/0.743

videos sampled from Flickr, while LIVE-VQC contains 585 10-second multiple-resolution videos captured by mobile devices. When defining the parametric temporal pooling models, we used $p = 2$ ($\mathcal{L}^2$) for Minkowski, $k = 10\%$ for percentile, $(L, \alpha_p, \alpha_r) = (180, 0.01, 0.01)$ for primacy and recency, and $(\tau, \alpha) = (60, 0.8)$ for Temporal Hysteresis, as recommended in the originating works. We randomly split the evaluation dataset into 80%-20% portions for training and testing, respectively, over 100 trials and report the overall median performance on the testing set. We only conducted 20 iterations for CORNIA due to its high training complexity. Within each split iteration, EPooling requires two phases of training – first, to train the mapping from the IQA feature vector to frame-level quality predictions, then, to learn the aggregation function that fuses the several pooled predictions to obtain a single quality score. We used a support vector regression (SVR) as the learning model for both training stages, employing cross-validation with 3×3 grid-search for the SVR parameter selection. As performance metrics, we used the Spearman rank-order correlation coefficient (SRCC) calculated between the ground truth MOS and the predicted scores to measure the prediction monotonicity of the models, and the Pearson linear correlation coefficient (PLCC) (computed after logistic mapping) to measure the degree of linear correlation against MOS.

4.2. Results and Recipe

The performance results are shown in Table 1 on the KoNViD-1k [34] and LIVE-VQC [35], respectively. On KoNViD-1k, none of the sophisticated pooling algorithms were observed to significantly outperform the sample mean of temporal video quality scores. While an average gain of ~ 0.01 in SRCC/PLCC was achieved using Hysteresis pooling, the three classical Pythagorean means performed quite well despite their simplicity and computational efficiency. When tested on LIVE-VQC [35], however, we have observed a ~ 0.03 average performance gain when employing perceptual importance pooling like percentile [14], VQPooling [16], and Hysteresis [19], regardless of which IQA model was used. It is likely that the memory-related effects, primacy and recency, would play a more important role on longer videos (usually minutes long), as shown in [18, 25], but they did not contribute much on the short duration videos (8-10 seconds) in these datasets. Our proposed ensemble method of pooling achieved consistently competitive outcomes on both datasets.

These performance results yet reveal different trends on the two databases: KoNViD-1k yielded similar results among most of the competing pooling approaches, whereas on LIVE-VQC, Percentile, VQPooling, Hysteresis, and the ensemble enhancement, EPooling, generated the best scores. Towards understanding this, we observe that LIVE-VQC contains videos with more camera motion, hence more temporal variation than those in KoNViD-1k. It is possible that LIVE-VQC contains a larger range of perceived time-varying qualities scores, while temporal quality variations in KoNViD-1k adhere more closely to the mean quality level. Recalling the aforementioned hypothesis that perceptual quality is heavily affected by the worst portions of a video, our experimental results promote this assumption. In conclusion, our suggested recipe for incorporating temporal pooling into the design of NR VQA models strongly depends on video content – for videos containing more motion or temporal quality variations, pooling strategies that more heavily weight low quality events are recommended. In situations where the quality variations are low, or contain less motion, traditional statistical mean predictions may be adequate.

5. CONCLUSION

We conducted a benchmark study on the added value of integrating temporal pooling into blind video quality assessment for user-generated video content. We found that the efficacy of temporal pooling is content-dependent, but an ensemble approach can further improve quality prediction performance on a difficult problem that is only incompletely understood.

6. REFERENCES

[1] 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, 2010.

[2] 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, 2004.

[3] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, 2016.

[4] 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, 2012.

[5] 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, 2014.

[6] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 491–495.

[7] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, 2015.

[8] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, pp. 32–32, 2017.

[9] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2351–2359.

[10] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *arXiv preprint arXiv:2005.14354*, 2020.

[11] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. Cock, "VMAF: The journey continues," *Netflix Technology Blog*, 2018.

[12] S. Rimac-Drlje, M. Vranjes, and D. Zagar, "Influence of temporal pooling method on the objective video quality evaluation," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, 2009, pp. 1–5.

[13] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, ""To pool or not to pool": a comparison of temporal pooling methods for http adaptive video streaming," in *Proc. Int. Workshop Qual. Multimedia Exper. (QoMEX)*, 2013, pp. 52–57.

[14] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, 2009.

[15] C. Chen, M. Izadi, and A. Kokaram, "A perceptual quality metric for videos distorted by spatially correlated noise," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1277–1285.

[16] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, 2012.

[17] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik, "Bband index: a no-reference banding artifact predictor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 2712–2716.

[18] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5217–5231, 2017.

[19] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 1153–1156.

[20] L. K. Choi and A. C. Bovik, "Video quality assessment accounting for temporal visual masking of local flicker," *Signal Process.: Image Commun.*, vol. 67, pp. 182–198, 2018.

[21] M. A. Aabed and G. AlRegib, "Perceptual video quality assessment: Spatiotemporal pooling strategies for different distortions and visual maps," in *Proc. IEEE Int. Workshop Multimedia Signal Process. (MMSP)*. IEEE, 2016, pp. 1–6.

[22] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, 2011.

[23] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, 2010.

[24] B. B. Murdock Jr, "The serial position effect of free recall," *J. Exp. Psychol.*, vol. 64, no. 5, p. 482, 1962.

[25] D. Ghadiyaram, J. Pan, and A. C. Bovik, "Learning a continuous-time streaming video QoE model," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2257–2271, 2018.

[26] T. Nguyen Duc, C. Minh Tran, P. X. Tan, and E. Kamioka, "Modeling of cumulative QoE in on-demand video services: Role of memory effect and degree of interest," *Future Internet*, vol. 11, no. 8, p. 171, 2019.

[27] 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, 2012.

[28] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, 2017.

[29] J. De Vriendt, D. De Vleeschauwer, and D. Robinson, "Model for estimating QoE of video delivered using HTTP adaptive streaming," in *Proc. IFIP/IEEE Int Symp. Integr. Netw. Manag.*, 2013, pp. 1288–1293.

[30] C. Chen, L. K. Choi, G. De Veciana, C. Caramanis, R. W. Heath, and A. C. Bovik, "Modeling the time-varying subjective quality of HTTP video streams with rate adaptations," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2206–2221, 2014.

[31] M. Mirkovic, P. Vrgovic, D. Culibrk, D. Stefanovic, and A. Anderla, "Evaluating the role of content in subjective video quality assessment," *Sci. World J.*, vol. 2014, 2014.

[32] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 253–265, 2009.

[33] H. G. Longbotham and A. C. Bovik, "Theory of order statistic filters and their relationship to linear fir filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 2, pp. 275–287, 1989.

[34] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," in *Proc. Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2017, pp. 1–6.

[35] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, 2018.

[36] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, 2018.

[37] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual DOG model fused with random forest," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3282–3292, 2015.

[38] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, 2014.

[39] 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, 2017.

[40] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1098–1105.