Quality-of-Experience for Adaptive Streaming Videos: An Expectation Confirmation Theory Motivated Approach

Zhengfang Duanmu, *Student Member, IEEE*, Kede Ma, *Member, IEEE*, and Zhou Wang, *Fellow, IEEE*

***Abstract*— The dynamic adaptive streaming over HTTP provides an inter-operable solution to overcome volatile network conditions, but how the human visual quality of experience (QoE) changes with time-varying video quality is not well-understood. Here, we build a large-scale video database of time-varying quality and design a series of subjective experiments to investigate how humans respond to compression level, spatial and temporal resolution adaptations. Our path-analytic results show that quality adaptations influence the QoE by modifying the perceived quality of subsequent video segments. Specifically, the quality deviation introduced by quality adaptations is asymmetric with respect to the adaptation direction, which is further influenced by other factors such as compression level and content. Furthermore, we propose an objective QoE model by integrating the empirical findings from our subjective experiments and the expectation confirmation theory (ECT). Experimental results show that the proposed ECT-QoE model is in close agreement with subjective opinions and significantly outperforms existing QoE models. The video database together with the code is available online at https://ece.uwaterloo.ca/ zduanmu/tip2018ectqoe/.**

***Index Terms*— Quality-of-experience, video quality assessment, expectation confirmation theory.**

I. Introduction

THANKS to the fast development of network services and the remarkable popularity of smart mobile devices in the past decade, there has been a tremendous growth in streaming media applications, especially the wide usage of the dynamic adaptive streaming schemes over HTTP (DASH). Aiming to provide a good balance between the fluent experience and the video quality for better quality-of-experience (QoE), DASH video players adaptively switch among multiple available video streams of different bitrates, spatial resolutions, and frame rates based on various factors, including playback rate, buffer condition, and instantaneous throughput [2].

Manuscript received November 20, 2017; revised April 16, 2018 and May 28, 2018; accepted June 22, 2018. Date of publication July 12, 2018; date of current version September 17, 2018. This work was supported by the Natural Sciences and Engineering Research Council of Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dacheng Tao. This paper was presented in part at the 25th ACM Multimedia, Mountain View, CA, USA, October 2017 [1]. *(Corresponding author: Zhengfang Duanmu.)*

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: zduanmu@uwaterloo.ca; k29ma@uwaterloo.ca; zhou.wang@uwaterloo.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2855403

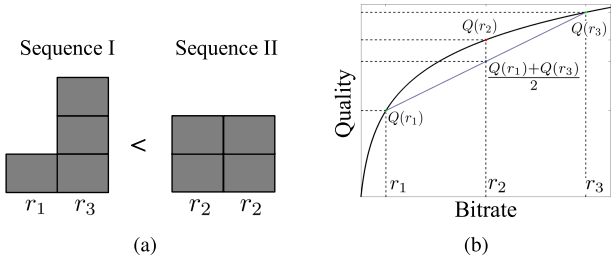


Fig. 1. Constant bitrate contour experiment fails to differentiate the effect of quality adaptations and the overall intrinsic quality of multiple video segments. Each rectangle in a column represents a fixed bitrate. (a) Constant bitrate contour test case 1. (b) Rate-quality curve.

Despite the widespread deployment of adaptive streaming technologies, our understanding of human QoE behaviors in this multi-dimensional adaptation space remains rather limited. Traditional adaptive bitrate selection algorithms ignore the impact of quality adaptations [3], which may result in unnecessary stalling events and lead to strong degradations in QoE [4]. To make the best use of adaptive streaming technologies, it is important to thoroughly understand the impact of quality adaptations on end-users' QoE.

Since the human visual system (HVS) is the ultimate receiver of streaming videos, subjective evaluation is the most straightforward and reliable approach to evaluate the QoE. Traditional subjective experiments investigate the impact of quality adaptations by varying the temporal video bitrate distributions in a constant average bitrate contour as illustrated in Fig. 1 and Fig. 2. Typical conclusions include 1) stronger adaptations lead to larger degradations in QoE and 2) users prefer positive over negative adaptations. However, this setup is problematic for two reasons. First, the HVS is complex and highly nonlinear. Perceptual quality generally is a concave function of the bitrate [5]. Therefore, a video sequence with a higher bitrate variance may have intrinsically lower average perceptual quality, regardless of quality adaptations. In Fig. 1, two video sequences have the same average bitrate but different temporal bitrate distributions. It is easy to show that the average perceptual quality of Sequence I $\frac{Q(r_1)+Q(r_3)}{2}$ is lower than that of Sequence II $Q(r_2)$. Similar conclusions can be drawn for other encoding configurations such as quantization parameter (QP), spatial resolution, and temporal resolution [6]. Second, video content at the end of a sequence tends to have a

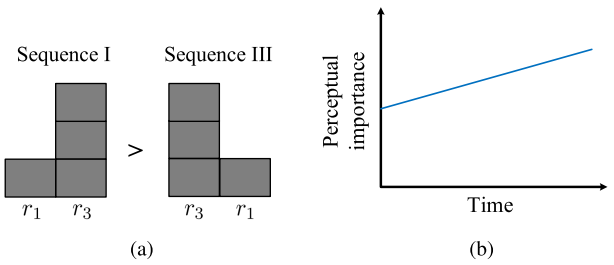


Fig. 2. Constant bitrate contour experiment confounds the effect of quality adaptation direction with the recency effect. (a) Constant bitrate contour test case 2. (b) Recency effect.

stronger impact on QoE, a phenomenon known as the recency effect [7]. Thus, the worse quality of Sequence III in Fig. 2 may be a consequence of the recency effect rather than the quality adaptation direction. We argue that both ambiguities perplex the conclusions drawn from existing subjective studies.

In this work, we first carry out three subjective experiments to address the aforementioned confounding factors and to better explore the space of quality adaptations. In Experiment I, we evaluate the quality of short video segments (four seconds, 4s) at various compression levels, spatial resolutions, and frame rates. In Experiment II, we concatenate 4s segments of the same content from Experiment I into long video sequences (eight seconds, 8s) to simulate quality adaptation events. Separate subjective opinions are collected for the two 4s segments after watching the whole sequence. In Experiment III, subjects provide a single score to reflect their overall QoE on the concatenated 8s video sequences. The subjective results suggest that quality adaptations alter the perceived quality of the second video segments, which consequently influence the overall QoE. Moreover, existing objective streaming video QoE models fail to accurately predict the QoE in our more realistic testing environment. This motivates us to develop a new framework for objective QoE based on the expectation confirmation theory (ECT) [8], which is widely applied in consumer behavior research but has not been exploited in the context of QoE prediction. We construct an ECT-based QoE measure (ECT-QoE), which takes spatial and temporal expectation confirmations into separate considerations. Experiments show that ECT-QoE significantly outperforms state-of-the-art objective models. In addition, ECT-QoE is instantaneous, making it ideal for the optimization of media streaming systems.

II. Related Work

A. Subjective QoE on Time-Varying Video Quality

A significant number of subjective QoE studies have been conducted to understand time-varying video quality. Two excellent surveys can be found in [9] and [10]. Here we only provide a brief overview.

Zink *et al.* [11], [12] made one of the first attempts to measure the perceptual experience of scalable videos of similar average peak signal-to-noise ratio (PSNR) with different variances. This constant contour experimental design was later adopted by other subjective studies [4], [13] to investigate the QoE of adaptive streaming videos. The conclusions drawn from such design are not well grounded, as discussed in Section I. To overcome the limitations of the constant contour strategy, a few subjective experiments [14], [15] investigated how subjects react to a video of significant time-varying quality. However, ambiguities remain on the influence of switching and the recency effect, as exemplified in Fig. 2. Moreover, the scope of the studies was limited to bitrate adaptations only.

Several other subjective studies [16]–[24] have been conducted without variable control, mainly towards identifying influencing factors of QoE and benchmarking adaptive bitrate selection algorithms. Although negative quality adaptations are commonly considered annoying, no agreement was reached upon how positive quality adaptations affect the QoE. Three contradictory theoretical positions have been put forth: positive adaptations introduce reward [4], [21], penalty [18], [19], [23], or no effect [22].

In addition, all existing studies suffer from one or more of the following limitations: (1) the datasets are very limited in size; (2) multi-dimensional adaptations that commonly occur in practice are not presented; and (3) most datasets are not publicly available for reproduction and further investigation.

B. Objective Models on Time-Varying Video Quality

Modern video quality assessment (VQA) models typically operate on local video frames/segments and combine local quality scores into a single scalar score. Despite the success in predicting the perceptual quality of statically encoded videos, this scheme often falls short of evaluating time-varying video quality. A significant number of algorithms have been proposed to overcome the limitation, which can be roughly categorized as follows.

1) Temporal Pooling: A common hypothesis of perceptual pooling schemes is that the relative importance of local video quality scores is correlated with visual attention. Memory-based pooling [25] was developed to account for the recency effect [7]. Based on the hypothesis that severe impairments have a substantial impact on human quality judgments, various studies adopted distortion-based strategies [26]–[30]. Somewhat surprisingly, none of sophisticated temporal pooling methods is shown to consistently outperform the average pooling in recent comparative studies [4], [14], [31].

2) Switching Experience Quantification: Several studies hypothesized that quality adaptations have a direct impact on the QoE of adaptive streaming videos, which can be modeled as a combination of intrinsic video quality and adaptation experience. Using either bitrate or objective VQA measurement as the base quality indicator, switching experience has been modeled through total variation [32], [33], variance [22], asymmetric piecewise linear function [14], random forest regression [34], Hammerstein-Wiener model [35], and nonlinear autoregressive model [36]. Nevertheless, these algorithms lack comprehensive verifications on databases that contain a good coverage of video content variations. Most existing objective time-varying VQA models focus on presentation quality adaptations only. State-of-the-art streaming algorithms, however, may adapt presentation quality, spatial resolution,

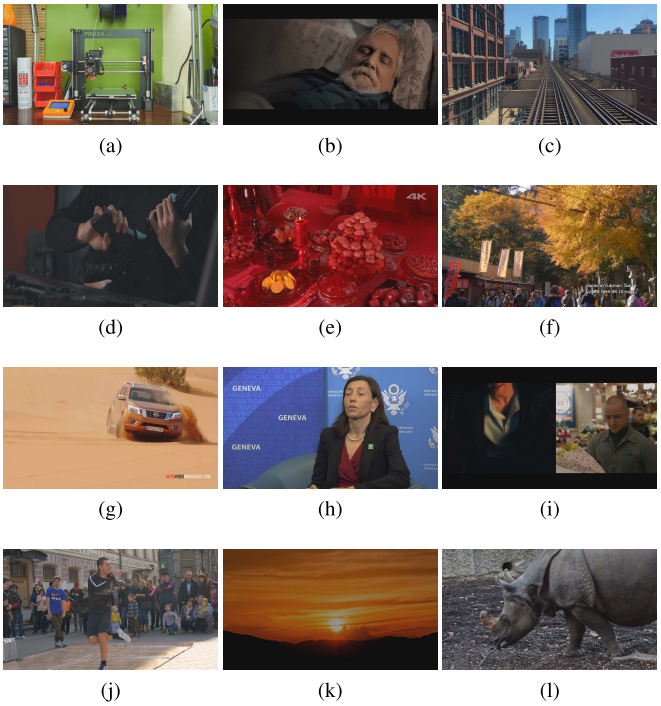


Fig. 3. Snapshots of reference video sequences. (a) 3dPrinter. (b) Armenchik. (c) Chicago. (d) FightForGlory. (e) Fruits. (f) MtTakao. (g) Navara. (h) News. (i) SplitTrailer. (j) StreetDance. (k) Sunrise. (l) WildAnimal.

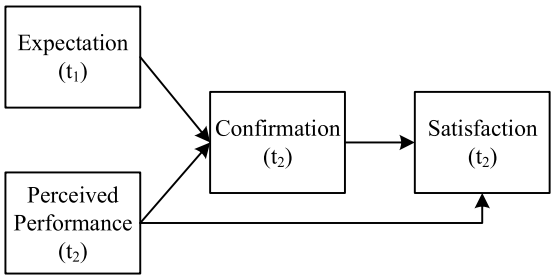


Fig. 4. General framework of expectation confirmation theory. t_1 and t_2 represent pre-consumption and post-consumption variables, respectively.

frame rate, or a combination of them at a certain time instance. The generalizability of the existing objective models in such a multi-dimensional adaptation space has not been systematically examined. More importantly, existing models tend to be ad-hoc, lacking connections to theoretic frameworks of human behaviors as the basis for developing reliable computation models.

C. Expectation Confirmation Theory (ECT)

ECT is widely used in the consumer behavior literature to study the consumer satisfaction [8]. The key constructs and relationships in ECT are illustrated in Fig. 4. Consumers form an initial expectation of a specific product or service prior to purchase. Following a period of initial consumption, they form perceptions about its performance. They assess its performance with respect to their original expectation and determine the extent to which their expectation is confirmed.

TABLE I

CHARACTERISTICS OF REFERENCE VIDEOS. SI: SPATIAL INFORMATION. TI: TEMPORAL INFORMATION. HIGHER SI/TI INDICATES HIGHER SPATIAL/TEMPORAL COMPLEXITY

Index	Name	SI	TI	Description
a	3dPrinter	81	67	Indoor scene, smooth motion
b	Armenchik	30	33	Human, smooth motion
c	Chicago	108	30	Architecture, high motion
d	FightForGlory	17	33	Human, average motion
e	Fruits	45	32	Plants, smooth motion
f	MtTakao	95	72	Natural scene, average motion
g	Navara	31	41	Transportation, smooth motion
h	News	76	8	News, smooth motion
i	SplitTrailer	58	48	Movie, smooth motion
j	StreetDance	58	19	Sport, high motion
k	Sunrise	64	56	Natural scene, high motion
l	WildAnimal	77	49	Animal, average motion

TABLE II

ENCODING LADDER OF VIDEO SEQUENCES. QP: QUANTIZATION PARAMETER. fps: frames/second

Representation	Resolution	QP	fps
Q_1	1920×1080	48	30
Q_2	1920×1080	≈ 40	30
Q_3	1920×1080	32	30
S_1	480×270	32	30
S_2	768×432	32	30
T_1	1920×1080	32	5
T_2	1920×1080	32	10

Finally, they form a satisfaction based on their confirmation level and performance.

The predictive ability of ECT has been demonstrated in a wide range of contexts, including automobile repurchase [37], camcorder repurchase [38], institutional repurchase of photographic products [39], restaurant services [40], business professional services [41], and information system continuance [42]. In Section IV, we extend the application scope of ECT in the context of user experience assessment in streaming media consumption.

III. DATABASE AND SUBJECTIVE EXPERIMENTS

A. Video Database Construction

We construct a new video database, which contains 12 pristine high-quality videos and spans diverse content, including humans, plants, natural scenes, news, and architectures. The detailed specifications are listed in Table I and the screenshots are shown in Fig. 3. Spatial information (SI) and temporal information (TI) [43] that roughly reflect the complexity of video content are also given in Table I, where larger SI/TI indicates higher spatial/temporal complexity. Apparently, the video sequences are of diverse spatio-temporal complexity. An 8s video sequence [44] is extracted from each source video, which is further partitioned into two non-overlapping 4s segments, referred to as short segments (SS). We encode each SS into seven representations using H.264 according to the encoding ladder shown in Table II. An internal subjective test is conducted to divide the seven representations into three sets $\{Q_1, S_1, T_1\}$, $\{Q_2, S_2, T_2\}$, and $\{Q_3\}$ corresponding to low-, medium-, and high-quality levels, respectively. To simulate quality adaptation events in adaptive streaming,

TABLE III

ADAPTATION TYPES. Q-Q: COMPRESSION LEVEL ADAPTATION. S-S: SPATIAL RESOLUTION ADAPTATION. T-T: TEMPORAL RESOLUTION ADAPTATION. Q-S: COMPRESSION LEVEL AND SPATIAL RESOLUTION ADAPTATION. Q-T: COMPRESSION LEVEL AND TEMPORAL RESOLUTION ADAPTATION. S-T: SPATIAL RESOLUTION AND TEMPORAL RESOLUTION ADAPTATION

Adaptation type	Adaptation intensity				
	ΔQ=-2	ΔQ=-1	ΔQ=0	ΔQ=1	ΔQ=2
Q-Q	Q_3Q_1	Q_2Q_1, Q_3Q_2	Q_1Q_1, Q_2Q_2, Q_3Q_3	Q_1Q_2, Q_2Q_3	Q_1Q_3
S-S	Q_3S_1	S_2S_1, Q_3S_2	S_1S_1, S_2S_2	S_1S_2, S_2Q_3	S_1Q_3
T-T	Q_3T_1	T_2T_1, Q_3T_2	T_1T_1, T_2T_2	T_1T_2, T_2Q_3	T_1Q_3
Q-S	–	Q_2S_1, S_2Q_1	Q_1S_1, S_1Q_1, Q_2S_2, S_2Q_2	Q_1S_2, S_1Q_2	–
Q-T	–	Q_2T_1, T_2Q_1	Q_1T_1, T_1Q_1, Q_2T_2, T_2Q_2	Q_1T_2, T_1Q_2	–
S-T	–	S_2T_1, T_2S_1	S_1T_1, T_1S_1, S_2T_2, T_2S_2	S_1T_2, T_1S_2	–

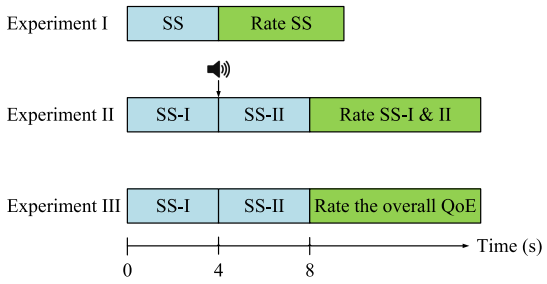


Fig. 5. Experimental procedures. SS: short segments include both SS-I and SS-II representing the first 4-second and last 4-second videos, respectively.

we concatenate two consecutive 4s segments with different representations from the same content into an 8s long sequence (LS). Table III lists the quality adaptation patterns, from which we observe diverse adaptation intensities and types. Furthermore, to better exploit the space of adaptations, three multi-dimensional adaptations (Q-S, Q-T, and S-T) are also included. As a result, there are 168 4s SS and 588 8s LS in the database.

B. Subjective User Study

Our subjective experiments generally follow the absolute category rating methodology, as suggested by the ITU-T recommendation P.910 [43]. We carry out three subjective experiments as illustrated in Fig. 5. Subjects are invited to rate the quality of SS in Experiment I. The subjective mean opinion score (MOS) of each SS is referred to as the intrinsic quality. Experiment II is conducted on LS, wherein subjects give separate opinions to the first and second 4s video segments (referred to as SS-I and SS-II, respectively). An audio stimulus is introduced in the middle of each LS, indicating the end of SS-I and the start of SS-II. The audio stimulus is short and undisruptive in order not to interfere subjects' viewing experience. In Experiment III, subjects are requested to watch the LS and to provide a single score to reflect their overall QoE. In order to remove any memory effect, we randomly shuffle content and adaptation patterns. A training session is performed before each experiment to familiarize subjects with typical distortion types and levels. We limit the length of each session up to 25 minutes to reduce the fatigue effect. Subjects score the quality of each video sequence according to the eleven-grade quality scale [43].

The subjective testing is setup in a normal indoor home setting with an ordinary illumination level. All videos are displayed at their actual pixel resolution on an LCD monitor with 1920 × 1080 pixel resolution and Truecolor (32 bits). The monitor is calibrated in accordance with the ITU-T BT.500 recommendations [45]. A customized graphical user interface is used to render the videos on the screen and to record subject ratings. A total of 36 naïve subjects, including 16 males and 20 females aged between 18 and 33, participate in the subjective experiments. The subject rejection procedure in [45] is used and one is removed from the experiment. Considerable agreement is observed among different subjects on the perceived quality of test videos for all three experiments.

C. Experiments I and II

The intrinsic quality of SS is compared to the post-hoc quality of SS in Experiment II to investigate the influence of quality adaptations. As illustrated in Fig. 6 and Fig. 7, quality adaptations have substantially different impacts on the perceptual quality of video segments before and after the switching. The MOSs of SS-I are highly consistent in both experiments. However, adaptations change the subjects' strategy in updating their opinions on SS-II. Specifically, we identify four influencing factors of such quality deviations from intrinsic quality and summarize the observations as follows.

1) Intensity Effect: The quality intensity change is the dominant factor of the quality deviation of SS-II. Fig. 7 (a) shows that the perceptual quality of SS-II following a negative quality adaptation is generally lower than its intrinsic quality, and the amount of penalty is correlated with the intensity of the negative adaptation. One explanation may be that there is a higher expectation when viewers are exposed to high-quality video content in the beginning, and thus the quality degradation makes them feel more frustrated. The overall trend aligns with existing studies of time-varying video quality [4], [12], [13], [16], [20]. However, we do not observe a consistent penalty or reward for constant and positive quality adaptation scenarios.

2) Type Effect: The adaptation type, given in Table III, is another major influential factor of QoE. Significant differences between subjective ratings given to different adaptation types can be found in Fig. 7 (b). In particular, the temporal resolution adaptation is rated as the least favorable approach, even in the positive adaptation case. Compression level and

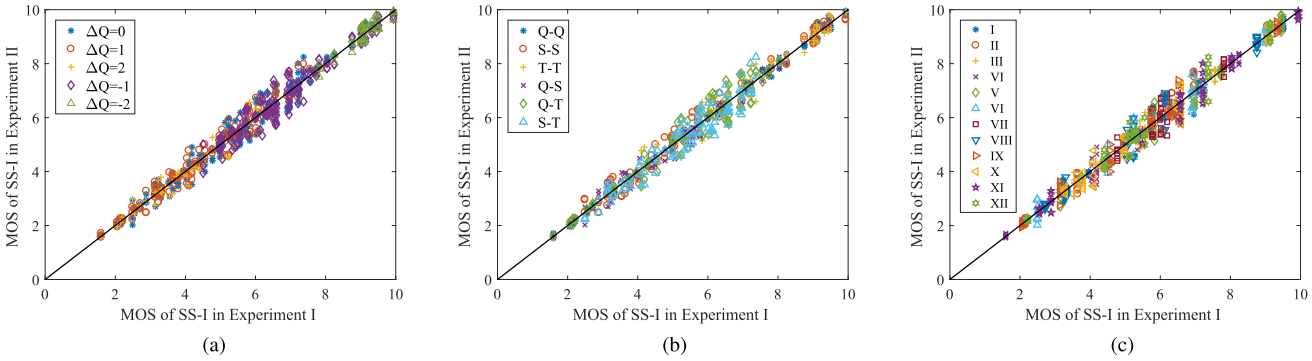


Fig. 6. MOS of SS-I in Experiment I vs. MOS of SS-I in Experiment II. (a) Intensity Effect. (b) Type Effect. (c) Content Effect.

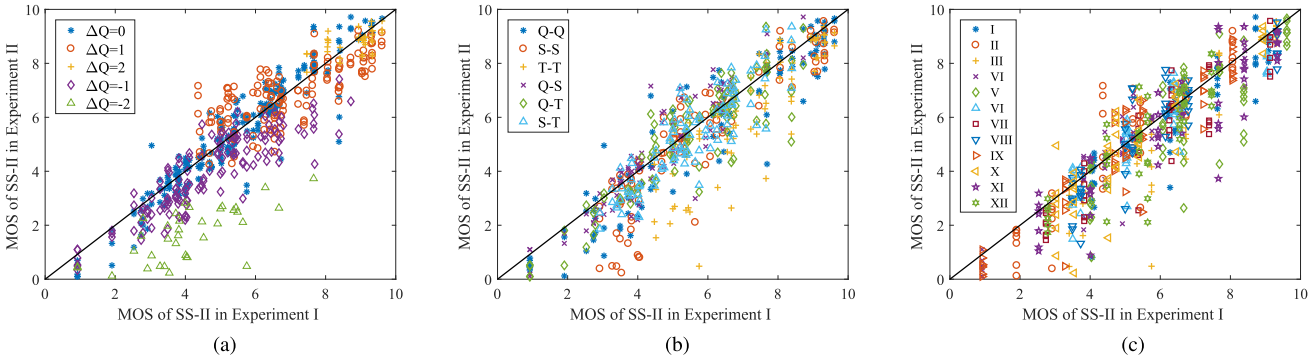


Fig. 7. MOS of SS-II in Experiment I vs. MOS of SS-II in Experiment II. (a) Intensity Effect. (b) Type Effect. (c) Content Effect.

spatial resolution adaptations do not introduce extra penalty in general, whereas subjects penalize sudden occurrence of blurring artifacts when the quality of SS-I is high. In addition, the multi-dimensional adaptation types Q-T and S-T introduce penalty on SS-II, especially when the intrinsic quality ranges from medium to high, while the Q-S adaptations do not have such effect.

3) Level Effect: The amount of reward or penalty that subjects give to SS-II is not only affected by the intensity and type effects, but also by the intrinsic quality where the adaptation occurs. The vertical distance from the triangle points to the diagonal line in Fig. 7 (a) increases along the horizontal axis, suggesting that a quality degradation occurred in high-quality has more impacts on QoE than one occurred in the low-quality range. Conversely, subjects tend to give high reward to quality improvement occurred in the low-quality range, suggesting an interesting Weber's law effect [46]. Nevertheless, the amount of reward is relatively small, indicating that subjects use asymmetric strategies in updating their opinions. To the best of our knowledge, this level effect has not been reported in the literature, and may explain the lack of consistency in quality adaptation results.

4) Content Effect: Video content seems to play a minor role in quality adaptations. Nevertheless, we observe that sequences without scene changes such as Chicago and StreetDance are more heavily degraded by quality adaptations than sequences of frequent scene changes such as 3dPrinter and Sunrise. This may be because the quality adaptations occurred within the same scene are more perceivable. This phenomenon is also orally confirmed by the participants at the end of their test sessions.

We further perform the analysis of variance (ANOVA) test on the MOSs of SS-II to understand the statistical significance of the influencing factors, where the p-value is set to 0.05. The results suggest that the adaptation intensity, adaptation type, intrinsic quality, content variation, the interactions between them are statistically significant to the MOS discrepancy between Experiments I and II.

D. Experiment III

To understand the strategy that subjects employed to integrate segment-level perceptual video quality into an overall QoE score, we evaluate five temporal pooling strategies using both the intrinsic quality and post-hoc quality obtained in Experiments I and II, respectively [14]. These include 1) average: $\mathbf{w}_1 = [1/2, 1/2]$, 2) early dominance: $\mathbf{w}_2 = [1, 0]$, 3) late dominance: $\mathbf{w}_3 = [0, 1]$, 4) increasing weights: $\mathbf{w}_4 = [1/3, 2/3]$, and 5) decreasing weights: $\mathbf{w}_5 = [2/3, 1/3]$. Spearman's rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC) [47], and perceptually weighted rank correlation index (PWRC) [48] between the predicted and ground-truth QoE scores are calculated in Tables IV, V, and VI, respectively. Average pooling of post-hoc segment-level scores exhibits the highest correlation, even outperforming the increasing weights pooling strategy that is designed to account for the recency effect. This suggests that the short-rm recency effect (in the scale of seconds) has only marginal impact in adaptive streaming. Furthermore, Fig. 8

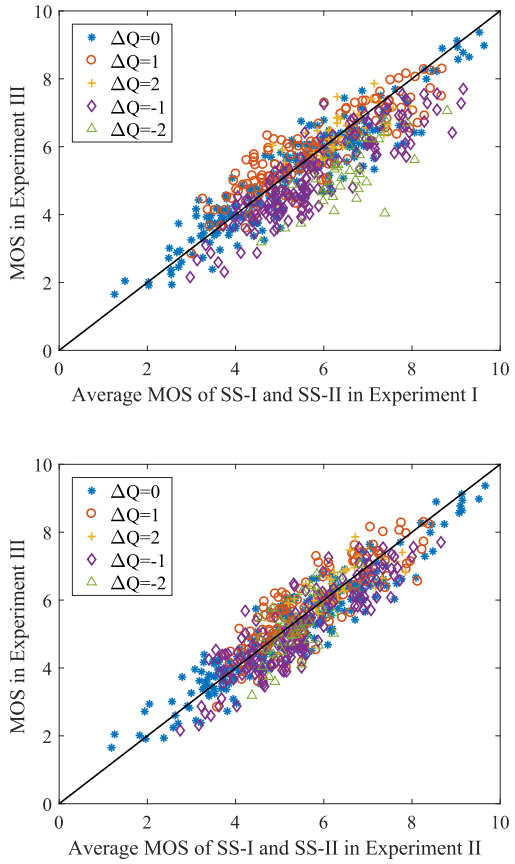


Fig. 8. Average pooling performance. (a) Experiment I. (b) Experiment II.

compares the scatter plots of the MOSs versus the average segment-level intrinsic quality scores and the average post-hoc quality scores, respectively. The average intrinsic quality tends to overestimate the QoE of LS when the negative quality adaptations are presented, while the average post-hoc quality achieves better performance. This observation suggests a promising approach in developing objective QoE models: instead of applying sophisticated temporal pooling strategies, we may first predict the segment-level post-hoc quality and then average pool the post-hoc quality scores as an estimation of the overall QoE.

E. Performance of Objective VQA Models

We test six objective VQA models including PSNR, SSIM [49], MS-SSIM [50], SSIMplus [51], VQM [52], and NIQE [53] along with five temporal pooling strategies as described in Section III-D. For each objective VQA algorithm, we average frame-level scores for each SS, resulting in 168 predicted scores. We then apply temporal pooling schemes on the segment-level scores for each LS. Since none of the full-reference VQA algorithms except for SSIMplus supports cross-resolution and cross-frame rate video quality evaluation, we up-sample all representations to 1920 × 1080 and 30 fps before evaluation. Tables IV, V, and VI summarize the results, which are somewhat disappointing because state-of-the-art VQA models and temporal pooling schemes provide moderately accurate predictions of time-varying video quality.

The test results also provide some useful insights regarding the general approaches used in VQA models. First, advanced VQA models, including SSIM, MS-SSIM, SSIMplus, and VQM, all significantly outperform the traditional PSNR measure, despite the fact that PSNR is still widely used in video encoding and streaming optimization approaches. SSIMplus, when combined with the increasing weight pooling, performs the best. Second, the straw-man solution of cross-frame rate VQA generally underestimates the quality of low frame rate video segments, suggesting that cross-frame rate VQA is a complex problem that requires more sophisticated modeling than what has been done in traditional VQA models. Third, none of the existing pooling strategies outperforms average pooling consistently.

F. Limitations and Extensions

Our conclusions on the influence of quality adaptations apply only to videos consisting of two segments, and its generalizability to multiple segments is an open question that is worthy further investigation. Nevertheless, our subjective experiment methodology makes it possible to better understand the influence of temporal dynamics in the QoE at the expense of reasonably higher workloads. The full set of experiments is approximately three times longer than the traditional single-stimulus methodology. Fortunately, with the development of static VQA algorithms, we expect the subjective evaluation on intrinsic video quality can be replaced by objective VQA algorithms, reducing the time complexity by a factor of 1.5. Thus, it is feasible to extend the experiment to accommodate longer video sequences by reducing the number of hypothetical reference circuits or source video sequences.

IV. Objective Quality Assessment

In this section, we propose a new framework for predicting streaming video QoE based on ECT. A brief introduction of ECT has been given in Section II-C, and here we focus on extending it to handle QoE prediction of time-varying video quality.

A. The ECT Framework for Time-Varying Video Quality

Our subjective results on post-hoc quality are conceptually in close agreement with ECT. Specifically, the post-hoc quality of SS-II depends on its intrinsic quality and the intensity of quality changes. Serving as the expectation of SS-II, SS-I provides the reference level for viewers to form judgments about the focal product. Such formulation of expectation may be generalized to longer video sequences by incorporating the self-perception theory [54], which posits that individuals continually adjust their expectation as they acquire new information about the focal behavior. The adjusted perceptions then provide the basis for subsequent behaviors. Thus, viewers' expectation on the quality of the n-th video segment can be modeled by the intrinsic quality of the $(n-1)$-th video segment, $Q_i(n-1)$. After the viewers watch the n-th video

TABLE IV

SRCC RESULTS USING DIFFERENT BASE QUALITY MEASURES (SEGMENT-LEVEL MOS, PSNR, SSIM, MS-SSIM, SSIMPLUS, VQM, NIQE, AND POST-HOC MOS) AND DIFFERENT POOLING STRATEGIES (AVERAGE, EARLY DOMINANCE, LATE DOMINANCE, INCREASING WEIGHTS, DECREASING WEIGHTS, AA [14], AND ECT-QOE). SEGMENT-LEVEL OBJECTIVE VQA IS COMPUTED AS FRAME AVERAGE

Base measure	PSNR	SSIM [49]	MS-SSIM [50]	SSIMplus [51]	VQM [52]	NIQE [53]	Segment-level MOS	*Post-hoc* MOS
Average	0.30	0.47	0.42	0.70	0.59	0.30	0.85	**0.90**
Early dominance	0.16	0.29	0.27	0.32	0.32	0.21	0.42	0.42
Late dominance	**0.37**	**0.58**	**0.55**	0.65	**0.64**	0.34	0.78	0.68
Increasing weights	0.33	0.50	0.45	**0.75**	0.62	0.32	**0.87**	0.86
Decreasing weights	0.25	0.43	0.38	0.59	0.51	0.28	0.72	0.76
AA [14]	0.28	0.49	**0.42**	0.66	0.56	0.33	0.83	–
ECT-QoE	**0.50**	**0.70**	**0.68**	**0.81**	**0.73**	**0.35**	**0.90**	**–**

TABLE V

PLCC RESULTS USING DIFFERENT BASE QUALITY MEASURES AND POOLING STRATEGIES

Base measure	PSNR	SSIM [49]	MS-SSIM [50]	SSIMplus [51]	VQM [52]	NIQE [53]	Segment-level MOS	Post-hoc MOS
Average	0.32	0.56	0.49	0.73	0.59	0.34	0.87	**0.90**
Early dominance	0.18	0.36	0.29	0.36	0.33	0.24	0.45	0.45
Late dominance	**0.39**	**0.61**	**0.56**	0.68	**0.62**	0.37	0.79	0.69
Increasing weights	0.35	0.59	0.53	**0.77**	**0.62**	0.35	**0.88**	0.87
Decreasing weights	0.28	0.50	0.44	0.62	0.52	0.32	0.75	0.79
AA [14]	0.30	0.56	0.49	0.68	0.57	0.35	0.82	–
ECT-QoE	**0.51**	**0.73**	**0.71**	**0.82**	**0.73**	**0.38**	**0.90**	**–**

TABLE VI

PWRC [48] RESULTS USING DIFFERENT BASE QUALITY MEASURES AND POOLING STRATEGIES

Base measure	PSNR	SSIM [49]	MS-SSIM [50]	SSIMplus [51]	VQM [52]	NIQE [53]	Segment-level MOS	*Post-hoc* MOS
Average	1.03	1.80	1.58	2.89	2.35	0.93	3.69	**3.93**
Early dominance	0.77	1.27	1.22	1.33	0.81	0.81	1.77	1.52
Late dominance	1.63	**2.46**	**2.32**	2.72	**2.68**	1.31	3.28	2.68
Increasing weights	1.18	1.96	1.73	**3.13**	2.49	1.00	**3.83**	3.74
Decreasing weights	0.83	1.62	1.40	2.31	1.96	0.81	2.95	3.19
AA [14]	0.84	1.89	1.58	2.55	2.09	0.92	3.33	–
ECT-QoE	**1.75**	**2.76**	**2.68**	**3.17**	**2.82**	**1.33**	**3.71**	**–**

segment, they evaluate the instantaneous QoE by comparing the intrinsic quality of the n-th segment with the previous viewing experience. Support of this association comes from the Helson's adaptation level theory [55], which postulates that humans perceive stimuli as a deviation from a baseline stimulus level. In the spirit of ECT, the confirmation of the n-th segment is formulated as $f(Q_i(n)-Q_i(n-1))$, where f is typically a nonlinear asymmetric function that mimics the human perceptual system, and the difference $Q_i(n)-Q_i(n-1)$ is a cognitive comparison between anticipated and received video quality. According to the traditional ECT [8], the asymmetric QoE response is a consequence of the joint effect of the General Negativity Theory [56] and the Contrast Theory [55]. The General Negativity Theory suggests that an unconfirmed expectation creates a state of psychological discomfort, because the actual performance contradicts the consumer's original hypothesis. As a result, both positive and negative quality adaptations induce a hedonically negative state in QoE. Contrast Theory assumes that the surprise of an unexpected stimulus results in exaggeration of the disparity between expected and actual stimulus properties, suggesting an extra reward/penalty. Consequently, subjects employ asymmetric strategies in evaluating the positive and negative quality adaptations. The instantaneous satisfaction, i.e., the post-hoc quality of the n-th segment $Q_p(n)$ is estimated as the summation of expectation and discrepancy perceptions

$$Q_p(n) = f\left(Q_i(n) - Q_i(n-1)\right) + Q_i(n), \tag{1}$$

where the additive relationship is assumed in the original ECT. In practice, one usually requires a single end-of-process QoE measure. From our subjective experiments, we know that the average post-hoc quality is an excellent indicator of the overall QoE

$$Q = \sum_{n=1}^{N} Q_p(n), \tag{2}$$

where N represents the total number of video segments.

Direct use of ECT, however, is not sufficient to capture the post-hoc quality, which is also influenced by the adaptation level and type effects. To be specific, ECT predicts the perceptual quality deviation as a function of adaptation intensity and does not take into consideration the interactions between the

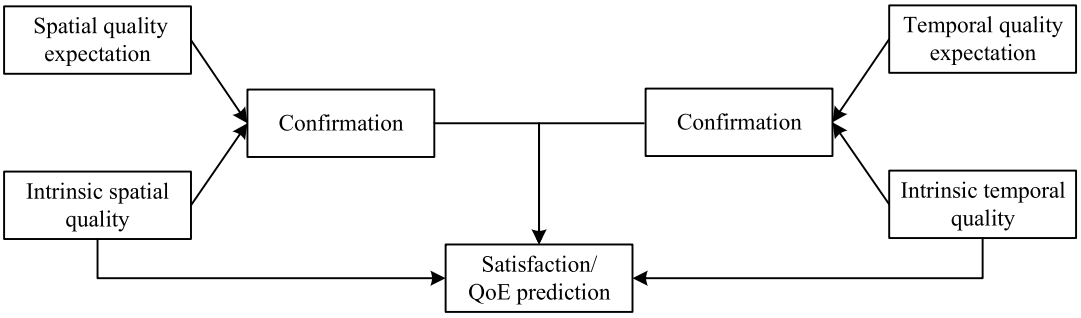


Fig. 9. The ECT framework for QoE prediction of time-varying video quality.

intrinsic quality and the quality adaptation intensity. To incorporate the level effect into ECT, we generalize Eq. (1) to

$$Q_p(n) = f\left(Q_i(n) - Q_i(n-1), Q_i(n)\right). \tag{3}$$

Another important fact we have learned from our subjective experiments is that different adaptation types have drastically different impacts on the post-hoc quality. While Q-T and S-T are generally perceived worse compared to Q-Q, S-S, and Q-S, they outperform T-T given the same initial intrinsic quality and adaptation intensity. Recognizing that adaptation types can be clustered into spatial, spatio-temporal, and temporal adaptations according to the impact on post-hoc quality, we extend the basic ECT to a multi-dimensional expectation confirmation process as shown in Fig. 9. We decompose the overall QoE into spatial quality and temporal quality, which are compared with their respective expectations. The resulting confirmations together with the intrinsic quality are combined into the post-hoc quality:

$$Q_p(n) = f\Big(\mathbf{Q}_i^S(n) - \mathbf{Q}_i^S(n-1), \mathbf{Q}_i^S(n), \mathbf{Q}_i^T(n) - \mathbf{Q}_i^T(n-1), \mathbf{Q}_i^T(n)\Big), \tag{4}$$

where $\mathbf{Q}_i^S$ and $\mathbf{Q}_i^T$ represent the intrinsic spatial quality and intrinsic temporal quality feature representations, respectively.

B. The ECT-QoE Model

We provide an instantiation of the ECT-based QoE framework. Specifically, for streaming videos, their instantaneous intrinsic quality $Q_i(n)$ can be estimated at the server side by a VQA model before transmission. Instead of frame-level quality, we choose to work with segment-level quality by averaging frame-level quality for the following reasons. First, segment-level solutions are practical in adaptive video streaming techniques, where video streams are encoded into a variety of bitrates and broken into HTTP file segments. The segment-level quality scores may be embedded in the manifest file that describes the specifications or carried in the metadata of the video container [57] at the beginning of a streaming session [22], [58], [59]. The availability of global video specifications is essential to improve the QoE because state-of-the-art adaptive streaming algorithms look ahead for future information when selecting the next video segment [32]. By contrast, although frame-level solutions achieve a higher temporal precision, current adaptive streaming techniques do not support the transmission of frame-level scores to the client. While it is possible to deploy no-reference frame-level VQA algorithms [60]–[62] on the client side, they are not as accurate and efficient as full-reference VQA models [4], [36], [63]. Second, segment-level evaluation is in closer agreement with human perception. As it is explained in ECT [8], the expectation confirmation is built after a period of video consumption. Third, the same coding configuration is applied to each segment in adaptive streaming videos, which are roughly constant in terms of content and complexity.

Since the ground-truth spatial/temporal intrinsic quality and confirmation are not available, we work on feature domain where objective VQA and spatial resolution are selected as spatial intrinsic quality representations, and frame rate as temporal intrinsic quality representations. The expectation confirmations are implemented as the differences of feature representations between the current and previous video segments. In summary, the input to the overall QoE prediction function f in Eq. (4) is a six-dimensional vector. We learn f using random forest regression [64]. To obtain the appropriate hyper-parameters in the model, we randomly split the dataset into disjoint 60% training, 20% validation, and 20% test sets. The random split is repeated 1000 times and the median SRCC, PLCC, and PWRC results are reported.

C. Validations

1) QoE Prediction Using Segment-MOS: To the best of our knowledge, there is no other publicly available video database that contains all of intrinsic quality, post-hoc quality, and ground-truth QoE. We first test ECT-QoE on our database using the segment-MOS (ground truth intrinsic quality) obtained in Experiment I. The PLCC, SRCC, and PWRC between MOSs and the predicted QoE scores are given in Tables IV, V, and VI. It can be observed that ECT-QoE outperforms all existing pooling strategies. The superiority of ECT-QoE is also evident in the scatter plots of Fig. 10.

2) QoE Prediction Using Objective VQA Models: Since expectation is a subjective quantity, it may not be available in many practical streaming applications. We employ PSNR, SSIM [49], MS-SSIM [50], SSIMplus [51], VQM [52], and NIQE [60] as the base video quality measures. To unify the scales used by different VQA models, we adopt a logistic

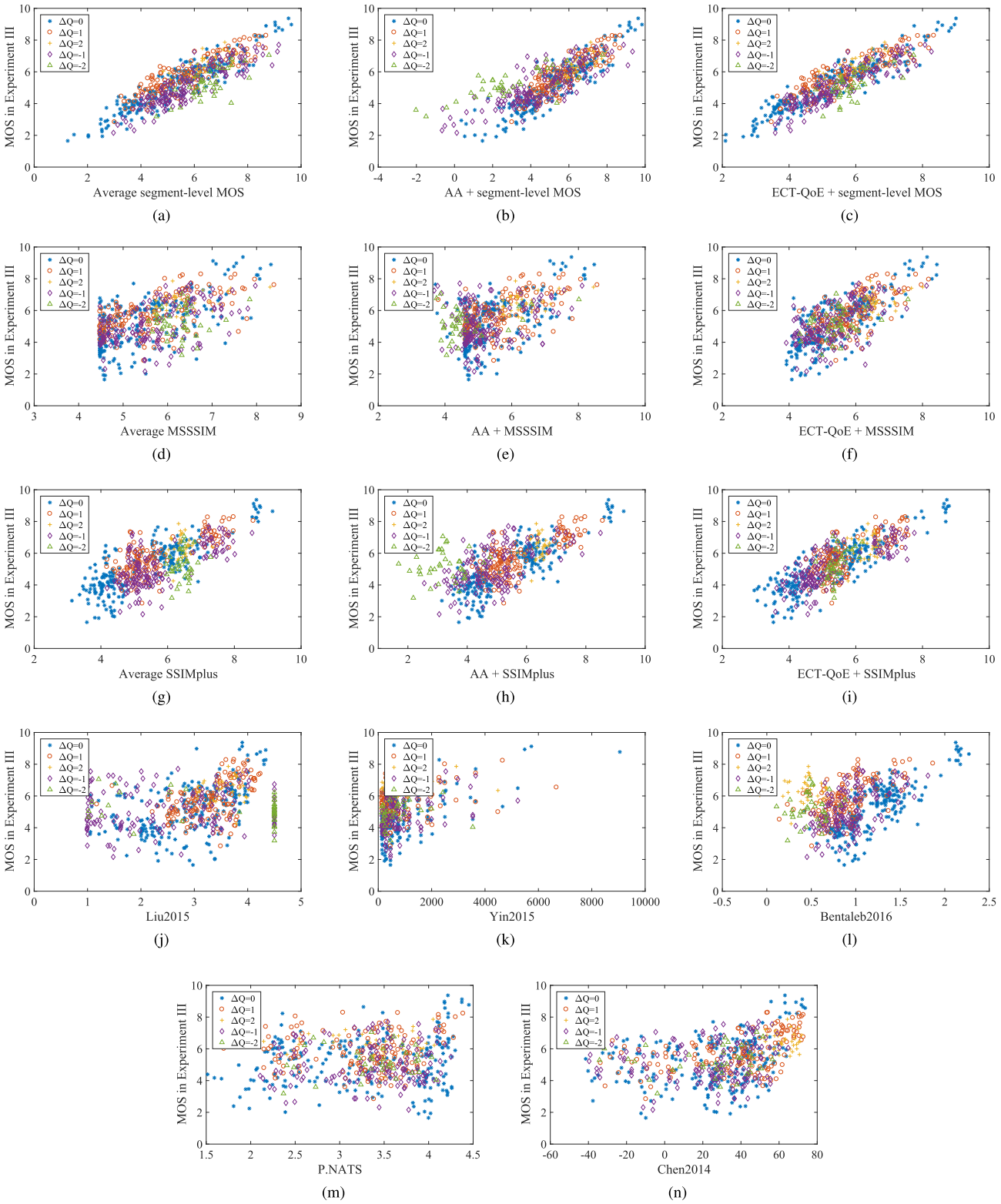


Fig. 10. Scatter plots of MOS in Experiment III versus prediction using different base quality measures and pooling strategies. (a) Average segment-level MOS. (b) Asymmetric adaptation (AA) [14] + segment-level MOS. (c) ECT-QoE + segment-level MOS. (d) Average MSSSIM. (e) AA + MSSSIM. (f) ECT-QoE + MSSSIM. (g) Average SSIMplus. (h) AA + SSIMplus. (i) ECT-QoE + SSIMplus. (j) Liu2015 [22]. (k) Yin2015 [32]. (l) Bentaleb2016 [33]. (m) P.NATS [34]. (n) Chen2014 [35].

nonlinear function as suggested in [47] to map the predictions of each model to the MOSs in Experiment I. The results are shown in Tables IV and V, where we observe that ECT-QoE significantly improves most base VQA methods. From the scatter plots in Fig. 10, we have two observations. First, by comparing each column, we can see that ECT-QoE

TABLE VII
COMPARISON WITH STATE-OF-THE-ART QOE MODELS

Model	SRCC	PLCC	PWRC
P.NATS [34]	0.07	0.08	0.30
Liu2015 [22]	0.37	0.41	1.22
Chen2014 [35]	0.39	0.35	1.37
Yin2015 [32]	0.42	0.42	1.43
Bentaleb2016 [33]	0.60	0.62	1.48
ECT-QoE	**0.81**	**0.82**	**3.71**

TABLE VIII
COMPARING ECT-QOE WITH ITS VARIANTS TO IDENTIFY THE ROLE OF INPUT FEATURES

Input features	SRCC	PLCC	PWRC
SSIMplus	0.75	0.74	3.10
SSIMplus + fps	0.79	0.75	3.12
ECT-QoE	**0.81**	**0.82**	**3.71**

TABLE IX
COMPARING ECT-QOE WITH ITS ECT REMOVED VARIANT

	SRCC	PLCC	PWRC
$\{Q_i(n)\}$	0.72	0.69	2.97
ECT-QoE	**0.81**	**0.82**	**3.71**

achieves a higher compactness in the scatter plots. Second, the best performance is obtained by combining ECT-QoE with SSIMplus [51].

Next, we compare ECT-QoE with existing QoE models including P.NATS [34], AA [14], Chen2014 [35], Rodríguez2014 [65], Liu2015 [22], Yin2015 [32], Bentaleb2016 [33], $\text{VsQM}_{\text{DASH}}$ [66], and NARX-QoE [36]. Unfortunately, the implementations of many models are not publicly available and the algorithms are not presented in sufficient details for reimplementation. Therefore, we are only able to implement the models whose parameters are stated in the original papers and can be evaluated on the current database. All models are tested using their default parameter settings. Table VII summarizes the results. Bentaleb2016 [33] that generally underestimates video quality with positive adaptations, AA accounts for the asymmetric strategies that subjects use in update their opinions and achieve slightly better performance. By decomposing the overall QoE into spatial and temporal expectation confirmation processes, ECT-QoE achieves the highest prediction accuracy.

3) Ablation Experiments: We conduct a series of ablation experiments to single out the core contributors of ECT-QoE. We first train the random forest regression model with different subsets of features to represent intrinsic quality. We show univariant and bi-variant regression models with the highest correlations in Table VIII. We observe that even without encoding configurations, ECT-QoE still outperforms the base quality measure SSIMplus. Moreover, frame rate modeling brings the prediction accuracy to the next stage. Further incorporating spatial resolution marginally improves the performance. We conclude that the expectation confirmation framework and the spatio-temporal adaptation interplay are the keys to the success of ECT-QoE.

To further analyze the impact of the expectation confirmation process, we construct a baseline by predicting the post-hoc quality with only SSIMplus, spatial resolution, and frame rate. The results are listed in Table IX, from which we see that predicting post-hoc quality without the past experience leads to inferior performance.

V. CONCLUSION AND DISCUSSION

We have presented a subjective experiment protocol to exploit the multi-dimensional adaptation space. Our path-analytic experimental results indicate that the perceptual quality deviation introduced by quality adaptations is a function of the adaptation intensity, adaptation type, intrinsic quality, content variation, and the interactions between them. By adapting and integrating ECT with theoretical and empirical findings from our subjective experiments, we theorize a new QoE framework for adaptive video streaming. The proposed framework is useful in better understanding the psychological behaviors of human subjects in evaluating time-varying video quality. We develop a practical and efficient instantiation, namely ECT-QoE, and show that it performs favorably against state-of-the-art methods. We find that using the intrinsic video quality of previous segment as the first-order approximation of "expectations" works well in practice. We wish our explorations in ECT will shed light on further research towards understanding and producing better models of QoE.

Many challenging problems remain to be solved. First, due to the limited capacity of the subjective experiments, we only investigate the impact of a single quality adaptation event to the QoE. A comprehensive study consisting of more content types and adaptation patterns is desired to better understand the behaviors of human viewers and to examine the generalizability of the current findings. Second, incorporating more advanced cross-resolution and cross-frame rate video quality assessment models have great potentials to further boost the performance. Third, the optimization of the existing video streaming frameworks based on the current findings is another challenging problem that desires further investigations.

REFERENCES

[1] Z. Duanmu, K. Ma, and Z. Wang, "Quality-of-experience of adaptive video streaming: Exploring the space of adaptations," in *Proc. ACM Int. Conf. Multimedia*, Mountain View, CA, USA, Oct. 2017, pp. 1752–1760.
[2] T. Stockhammer, "Dynamic adaptive streaming over HTTP: Standards and design principles," in *Proc. ACM Conf. Multimedia Syst.*, San Jose, CA, USA, Feb. 2011, pp. 133–144.
[3] DASH Industry Forum. (2013). *For Promotion of MPEG-DASH*. [Online]. Available: http://dashif.org
[4] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012.
[5] Z. Wang and A. C. Bovik, "Modern image quality assessment," in *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1. San Rafael, CA, USA: Morgan & Claypool, Jan. 2006, pp. 1–156.
[6] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2473–2486, Jun. 2014.

[7] D. S. Hands and S. E. Avons, "Recency and duration neglect in subjective assessment of television picture quality," *Appl. Cogn. Psychol.*, vol. 15, no. 6, pp. 639–657, Nov. 2001.

[8] R. L. Oliver, "A cognitive model of the antecedents and consequences of satisfaction decisions," *J. Marketing Res.*, vol. 17, no. 4, pp. 460–469, Nov. 1980.

[9] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 1st Quart., 2014.

[10] M.-N. Garcia *et al.*, "Quality of experience and HTTP adaptive streaming: A review of subjective studies," in *Proc. IEEE Int. Conf. Qual. Multimedia Exper.*, Singapore, Sep. 2014, pp. 141–146.

[11] M. Zink, O. Künzel, J. Schmitt, and R. Steinmetz, "Subjective impression of variations in layer encoded videos," in *Proc. IEEE/ACM Int. Workshop Qual. Service*, Berlin, Germany, Jun. 2003, pp. 137–154.

[12] M. Zink, J. Schmitt, and R. Steinmetz, "Layer-encoded video in scalable adaptive streaming," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 75–84, Feb. 2005.

[13] L. Yitong, S. Yun, M. Yinian, L. Jing, L. Qi, and Y. Dacheng, "A study on quality of experience for adaptive streaming service," in *Proc. IEEE Int. Conf. Commun. Workshop*, Budapest, Hungary, Jun. 2013, pp. 682–686.

[14] A. Rehman and Z. Wang, "Perceptual experience of time-varying video quality," in *Proc. IEEE Int. Conf. Qual. Multimedia Exper.*, Klagenfurt, Austria, Jul. 2013, pp. 218–223.

[15] J. V. Talens-Noguera, W. Zhang, and H. Liu, "Studying human behavioural responses to time-varying distortions for video quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Quebec City, QC, Canada, Sep. 2015, pp. 651–655.

[16] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen, "Flicker effects in adaptive video streaming to handheld devices," in *Proc. ACM Int. Conf. Multimedia*, Scottsdale, AR, USA, Nov./Dec. 2011, pp. 463–472.

[17] S. Tavakoli, K. Brunnström, K. Wang, B. Andrén, M. Shahid, and N. Garcia, "Subjective quality assessment of an adaptive video streaming model," *Proc. SPIE*, vol. 9016, pp. 90160K-1–90160K-13, Feb. 2014.

[18] D. C. Robinson, Y. Jutras, and V. Craciun, "Subjective video quality assessment of HTTP adaptive streaming technologies," *Bell Labs Tech. J.*, vol. 16, no. 4, pp. 5–23, Mar. 2012.

[19] B. J. Villa, K. De Moor, P. E. Heegaard, and A. Instefjord, "Investigating quality of experience in the context of adaptive video streaming: Findings from an experimental user study," in *Proc. Norsk Informatikkonferanse*, Stavanger, Norway, Nov. 2013, pp. 122–133.

[20] R. K. P. Mok, E. W. W. Chan, and R. K. C. Chang, "Measuring the quality of experience of HTTP video streaming," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage.*, Dublin, Ireland, May 2011, pp. 485–492.

[21] M. Grafl and C. Timmerer, "Representation switch smoothing for adaptive HTTP streaming," in *Proc. IEEE Int. Workshop Perceptual Qual. Syst.*, Vienna, Austria, Sep. 2013, pp. 178–183.

[22] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and validating user experience model for DASH video streaming," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 651–665, Dec. 2015.

[23] B. Lewcio, B. Belmudez, A. Mehmood, M. Wältermann, and S. Möller, "Video quality in next generation mobile networks—Perception of time-varying transmission," in *Proc. IEEE Int. Workshop Tech. Committee Commun. Qual. Rel.*, Naples, FL, USA, May 2011, pp. 1–6.

[24] Z. Duanmu, A. Rehman, and Z. Wang, "A quality-of-experience database for adaptive video streaming," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 474–487, Jun. 2018.

[25] M. Barkowsky, B. Eskofier, R. Bitto, J. Bialkowski, and A. Kaup, "Perceptually motivated spatial and temporal integration of pixel based video quality measures," in *Proc. Welcome Mobile Content Qual. Exper.*, Vancouver, BC, Canada, Aug. 2007, pp. 1–7.

[26] 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, Apr. 2009.

[27] 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, Apr. 2009.

[28] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Spatio-temporal quality pooling accounting for transient severe impairments and egomotion," in *Proc. IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 2509–2512.

[29] K. Lee, J. Park, S. Lee, and A. C. Bovik, "Temporal pooling of video quality estimates using perceptual motion models," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, Sep. 2010, pp. 2493–2496.

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

[31] 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. IEEE Int. Conf. Qual. Multimedia Exper.*, Klagenfurt, Austria, Jul. 2013, pp. 52–57.

[32] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 325–338, Oct. 2015.

[33] A. Bentaleb, A. C. Begen, and R. Zimmermann, "SDNDASH: Improving QoE of HTTP adaptive streaming using software defined networking," in *Proc. ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands, Oct. 2016, pp. 1296–1305.

[34] W. Robitza, M.-N. Garcia, and A. Raake, "A modular HTTP adaptive streaming QoE model—Candidate for ITU-T P.1203 ('P.NATS')," in *Proc. IEEE Int. Conf. Qual. Multimedia Exper.*, Erfurt, Germany, May 2017, pp. 1–6.

[35] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, Jr., 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, May 2014.

[36] C. G. Bampis, Z. Li, and A. C. Bovik, "Continuous prediction of streaming video QoE using dynamic networks," *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 1083–1087, Jul. 2017.

[37] R. L. Oliver, "Cognitive, affective, and attribute bases of the satisfaction response," *J. Consum. Res.*, vol. 20, no. 3, pp. 418–430, Dec. 1993.

[38] R. A. Spreng, S. B. MacKenzie, and R. W. Olshavsky, "A reexamination of the determinants of consumer satisfaction," *J. Marketing*, vol. 60, no. 3, pp. 15–32, Jul. 1996.

[39] P. A. Dabholkar, C. D. Shepherd, and D. I. Thorpe, "A comprehensive framework for service quality: An investigation of critical conceptual and measurement issues through a longitudinal study," *J. Retailing*, vol. 76, no. 2, pp. 139–173, Jun. 2000.

[40] J. E. Swan and I. F. Trawick, "Disconfirmation of expectations and satisfaction with a retail service," *J. Retailing*, vol. 57, no. 3, pp. 49–67, Jan. 1981.

[41] P. G. Patterson, L. W. Johnson, and R. A. Spreng, "Modeling the determinants of customer satisfaction for business-to-business professional services," *J. Acad. Marketing Sci.*, vol. 25, no. 1, pp. 4–17, Dec. 1997.

[42] A. Bhattacherjee, "Understanding information systems continuance: An expectation-confirmation model," *MIS Quart.*, vol. 25, no. 3, pp. 351–370, Sep. 2001.

[43] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-R BT.910, International Telecommunication Union, Feb. 2009.

[44] P. Fröhlich, S. Egger, R. Schatz, M. Mühlegger, K. Masuch, and B. Gardlo, "QoE in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment?" in *Proc. IEEE Int. Conf. Qual. Multimedia Exper.*, Yarra Valley, VIC, Australia, Jul. 2012, pp. 242–247.

[45] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500-12, International Telecommunication Union, Jan. 2012.

[46] G. Fechner, *Elements of Psychophysics*. New York, NY, USA: Breitkopf & Härtel, 1966.

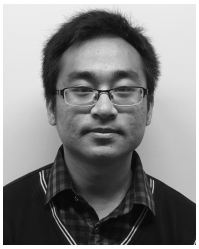
[47] VQEG. (2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment*. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

[48] Q. Wu, H. Li, F. Meng, and K. N. Ngan, "A perceptually weighted rank correlation indicator for objective image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2499–2513, May 2018.

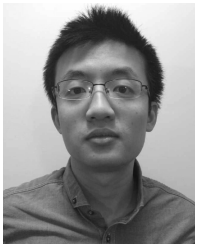
[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

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

[51] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," *Proc. SPIE*, vol. 9394, pp. 939406-1–939406-11, Mar. 2015.
[52] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.
[53] 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.
[54] D. J. Bem, "Self-perception theory," *Adv. Exp. Social Psychol.*, vol. 6, pp. 1–62, Jan. 1972.
[55] H. Helson, *Adaptation-Level Theory*. New York, NY, USA: Harper & Row, 1964.
[56] J. M. Carlsmith and E. Aronson, "Some hedonic consequences of the confirmation and disconfirmation of expectances," *J. Abnormal Social Psychol.*, vol. 66, no. 2, pp. 151–156, Feb. 1963.
[57] *Information Technology—MPEG Systems Technologies—Part 10: Carriage of Timed Metadata Metrics of Media in ISO Base Media File Format*, document ISO/IEC 23001-10, Sep. 2015. [Online]. Available: https://www.iso.org/standard/66064.html
[58] Z. Duanmu, A. Rehman, K. Zeng, and Z. Wang, "Quality-of-experience prediction for streaming video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Seattle, WA, USA, Jul. 2016, pp. 1–6.
[59] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 154–166, Feb. 2017.
[60] 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, Jan. 2016.
[61] Q. Wu, H. Li, F. Meng, and K. N. Ngan, "Toward a blind quality metric for temporally distorted streaming video," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 367–378, Jun. 2018.
[62] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.
[63] K. Ma *et al.*, "Group MAD competition? A new methodology to compare objective image quality models," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 1664–1673.
[64] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, Dec. 2002.
[65] D. Z. Rodríguez, Z. Wang, R. L. Rosa, and G. Bressan, "The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over HTTP," *EURASIP J. Wireless Commun. Netw.*, vol. 2014, no. 1, pp. 216–231, Dec. 2014.
[66] D. Z. Rodríguez, R. L. Rosa, E. C. Alfaia, J. I. Abrahão, and G. Bressan, "Video quality metric for streaming service using DASH standard," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 628–639, Sep. 2016.



Zhengfang Duanmu (S'15) received the B.A.Sc. and M.A.Sc. degrees from the University of Waterloo in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include perceptual image processing and quality of experience.



Kede Ma (S'13–M'18) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2014 and 2017, respectively. He is currently a Research Associate with the Howard Hughes Medical Institute and the Laboratory for Computational Vision, New York University, New York, NY, USA. His research interests include perceptual image processing, computational vision, and computational photography.

Zhou Wang (S'99–M'02–SM'12–F'14) received the Ph.D. degree from The University of Texas at Austin, Austin, TX, USA, in 2001. He is currently a Professor and the University Research Chair with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include image and video processing and coding, visual quality assessment and optimization, computational vision and pattern analysis, multimedia communications, and biomedical signal processing. He has over 200 publications in these fields with over 40 000 citations (Google Scholar).

Dr. Wang served as a member of the IEEE Multimedia Signal Processing Technical Committee from 2013 to 2015. He is a fellow of the Canadian Academy of Engineering. He was a recipient of the 2017 Faculty of Engineering Research Excellence Award at the University of Waterloo, the 2016 IEEE Signal Processing Society Sustained Impact Paper Award, the 2015 Primetime Engineering Emmy Award, the 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, the 2013 *IEEE Signal Processing Magazine* Best Paper Award, and the 2009 IEEE Signal Processing Society Best Paper Award. He has been serving as an Associate Editor for *Pattern Recognition* since 2006. He has been serving as a Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING since 2015 and an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY since 2016. He served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2009 to 2014 and the IEEE SIGNAL PROCESSING LETTERS from 2006 to 2010 and a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING from 2007 to 2009 and from 2013 to 2014.