Quality-of-Experience of Adaptive Video Streaming: Exploring the Space of Adaptations

Zhengfang Duanmu
University of Waterloo
zduanmu@uwaterloo.ca

Kede Ma
University of Waterloo
k29ma@uwaterloo.ca

Zhou Wang
University of Waterloo
Z.Wang@ece.uwaterloo.ca

ABSTRACT

With the remarkable growth of adaptive streaming media applications, especially the wide usage of dynamic adaptive streaming schemes over HTTP (DASH), it becomes ever more important to understand the perceptual quality-of-experience (QoE) of end users, who may be constantly experiencing adaptations (switchings) of video bitrate, spatial resolution, and frame-rate from one time segment to another in a scale of a few seconds. This is a sophisticated and challenging problem, for which existing visual studies provide very limited guidance. Here we build a new adaptive streaming video database and carry out a series of subjective experiments to understand human QoE behaviors in this multi-dimensional adaptation space. Our study leads to several useful findings. First, our path-analytic results show that quality deviation introduced by quality adaptation is asymmetric with respect to the adaptation direction (positive or negative), and is further influenced by the intensity of quality change (*intensity*), dimension of adaptation (*type*), intrinsic video quality (*level*), content, and the interactions between them. Second, we find that for the same intensity of quality adaptation, a positive adaptation occurred in the low-quality range has more impact on QoE, suggesting an interesting Weber's law effect; while such phenomenon is reversed for a negative adaptation. Third, existing objective video quality assessment models are very limited in predicting time-varying video quality.[1]

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Information systems** → **Multimedia streaming**;

KEYWORDS

Subjective video quality; Quality-of-experience (QoE) of end users; Adaptive video streaming; Layer switching

[1] The subjective database is available online at https://ece.uwaterloo.ca/~zduanmu/acmmm2017qoe/

MM'17, October 23–27, 2017, Mountain View, CA, USA

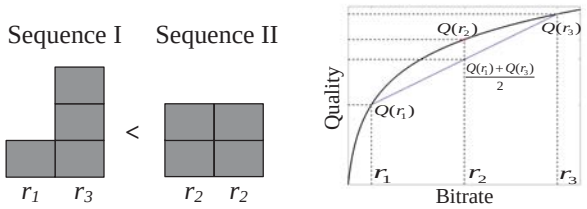
ACM ISBN 978-1-4503-4906-2/17/10...$15.00
https://doi.org/10.1145/3123266.3123418

1 INTRODUCTION

In the past decade, there has been a tremendous growth in streaming media applications, especially the wide usage of dynamic adaptive streaming schemes over HTTP (DASH), thanks to the fast development of network services and the remarkable growth of smart mobile devices. Aiming to provide a good balance between the fluent experience and the quality of videos for better quality-of-experience (QoE), DASH video player at the client adaptively switches among the available streams by varying video bitrate, spatial resolution, and frame-rate based on various factors, including playback rate, buffer condition, and instantaneous throughput [19].

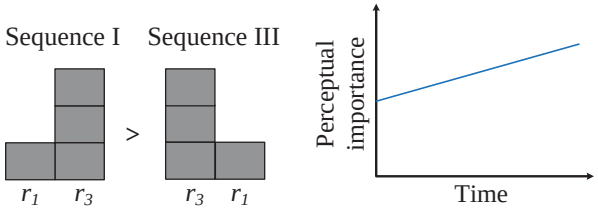
Despite the widespread deployment of adaptive streaming technology, our understanding of human QoE behaviors in this multi-dimensional adaptation space remains rather limited. Traditional adaptive bitrate selection algorithms ignore the impact of quality adaptations [1], leading to suboptimal performance. For example, incautious quality adaptation decisions may even result in a QoE as bad as stalling events [11]. To make efficient use of adaptive streaming technology, it is important to thoroughly understand the impact of quality adaptation in the 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 users' QoE of streaming videos. Traditional subjective experiments investigate the impact of quality adaptation 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) the correlation between the intensity of quality adaptation and the degradation in QoE and 2) the preference of positive over negative adaptations. However, this approach is problematic for two reasons. First, the HVS is complex and highly nonlinear. Perceptual quality generally exhibits a concave relationship with respect to the bitrate [25]. Therefore, it is unclear whether the lower rating of a video sequence with a higher bitrate variance is a consequence of quality adaptations or the lower average quality itself. For example, 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 $Q(r_2)$. Thus, the worse QoE of Sequence I does not necessarily suggest that subjects are annoyed by quality adaptation. Similar conclusions can be drawn for other encoding configurations such as quantization parameter (QP), spatial resolution, and temporal resolution [13]. Second, in subjective video quality assessment, scenes at the end of a sequence tend to have a



(a) Test case 1 (b) Typical rate-quality curve

Figure 1: Constant bitrate contour experiment fails to differentiate the effect of quality adaptation and the overall intrinsic quality of multiple video segments.



(a) Test case 2 (b) Recency effect

Figure 2: Constant bitrate contour experiment confounds the effect of quality adaptation direction with the recency effect.

stronger impact on QoE, a phenomenon known as the *recency effect* [6]. Thus, the worse quality of Sequence III in Fig. 2 may be a consequence of the *recency effect* rather than the quality adaptation direction. In summary, we argue that both ambiguities equivocate the validity of conclusions drawn from existing subjective studies.

In this work, we carry out three meticulously designed experiments to resolve the confounding factors and explore the space of quality adaptations. In Experiment I, we study the quality of short video segments (4-second) at various compression levels, spatial resolutions, and frame rates. In Experiment II, we concatenate the 4-second segments of the same content from Experiment I into long video sequences (8-second) to simulate quality adaptation events in adaptive streaming. Subjective user study is performed to collect separate opinions of the two short consecutive video segments after watching the whole video sequence. In Experiment III, subjects provide a single score to reflect their overall QoE on the concatenated video sequences. From the experimental results, we empirically show that quality adaptations alter the perceived quality of the second video segments, which consequently influences the overall QoE. More specifically, the intensity of quality change (*intensity*), dimension of adaptation (*type*), intrinsic video quality (*level*), content, and the interactions between them are influencing factors of perceptual quality deviation introduced by quality adaptations. Interestingly, we find that positive quality adaptations exhibit a Weber's law effect [2] similar to that observed in many other psychophysical test such as the perception of audio and visual contrast, but the effect is less significant for negative adaptations. Last, existing objective video quality assessment (VQA) models are very limited in predicting the time-varying video quality. All of these findings have significant implications on the future development of objective QoE models and QoE-driven adaptive video streaming schemes.

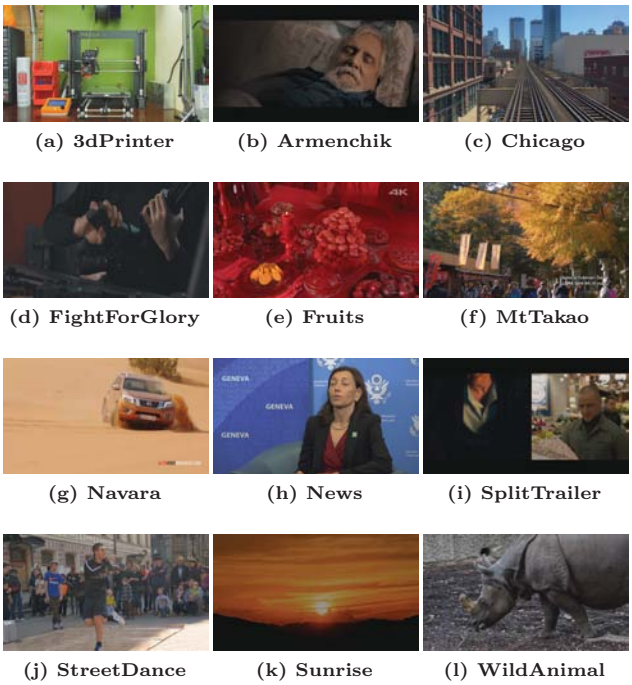
2 RELATED WORK

A significant number of subjective QoE studies have been conducted to understand the perceptual experience of time-varying video quality. Two excellent surveys on subjective QoE studies can be found in [18] and [4]. Here we only provide a brief overview.

Zink *et al.* [28, 29] made one of the first attempts to measure the perceptual experience of scalable videos. By investigating videos of similar average peak signal-to-noise ratio (PSNR) with different variances, they pointed out that the frequency and intensity of quality adaptations influence the perceived quality and should therefore be kept small. By experimenting with more video content, Moorthy *et al.* [11] and Liu *et al.* [9] came to the same conclusion that gradual quality variations are preferred. Moorthy *et al.* [11] confirmed the *recency effect* [6] in adaptive streaming, a phenomenon that can also described as "all is well if end is well" [4]. Ni *et al.* [12] investigated the influences of compression level, spatial resolution, and frame-rate adaptations on QoE of scalable coded videos. Besides similar conclusions, the authors also found that spatial and temporal resolution adaptations influence QoE in a content-dependent fashion. However, all the above mentioned subjective experiments based on constant bitrate contour design ignore the nonlinear perception of video quality.

To overcome the limitations of the constant encoding configuration contour design, Rehman *et al.* [15] and Talens-Noguera *et al.* [20] investigated how subjects react to a video consisting of multiple video clips that have significantly different perceptual quality. However, the authors still failed to answer whether quality adaptation itself affects QoE because the test procedure confounded the influence of switching and the *recency effect*, as exemplified in Fig. 2. Furthermore, the scope of the studies was limited to compression level adaptation.

Several other subjective studies [5, 7, 8, 10, 17, 21, 24] have been conducted without variable control, mainly towards identifying influencing factors of QoE and benchmarking adaptive bitrate selection algorithms. Interestingly, although many investigators agreed that negative quality adaptation is considered annoying, they do not agree upon how positive quality adaptation affects QoE. Three different theoretical positions have been put forth regarding to the influence of positive adaptation on QoE: positive adaptation introduces reward [5, 11], penalty [7, 17, 24], or no effect [8]. Furthermore,



(a) 3dPrinter (b) Armenchik (c) Chicago

(d) FightForGlory (e) Fruits (f) MtTakao

(g) Navara (h) News (i) SplitTrailer

(j) StreetDance (k) Sunrise (l) WildAnimal

Figure 3: Snapshots of reference video sequences.

Table 1: Characteristics of reference videos. SI: spatial information. TI: temporal information. Higher SI/TI indicates higher spatial/temporal complexity.

Index	Name	SI	TI	Description
I	3dPrinter	81	67	Indoor scene, smooth motion
II	Armenchik	30	33	Human, smooth motion
III	Chicago	108	30	Architecture, high motion
IV	FightForGlory	17	33	Human, average motion
V	Fruits	45	32	Plants, smooth motion
VI	MtTakao	95	72	Natural scene, average motion
VII	Navara	31	41	Transportation, smooth motion
VIII	News	76	8	News, smooth motion
IX	SplitTrailer	58	48	Movie, smooth motion
X	StreetDance	58	19	Sport, high motion
XI	Sunrise	64	56	Natural scene, high motion
XII	WildAnimal	77	49	Animal, average motion

the results of the experiments are not directly applicable to develop and validate computational models of time-varying video quality.

In summary, all of the above studies suffer from one or more of the following limitations: (1) the datasets are of insignificant size; (2) multi-dimensional adaptations commonly used in practice are not presented; and (3) the datasets are not available to the public (except for the LIVE Mobile VQA database [11]). Realizing the need for an adequate and more relevant resource, we construct a new publicly available database aiming for broader utility for modeling and analyzing time-varying video quality.

Table 2: Encoding ladder of video sequences

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

3 EXPERIMENT DESIGN

3.1 Video Database Construction

A video database of 12 pristine high-quality videos of size 1920×1080 and frame rate of 30 frames per second (fps) are selected to cover diverse content, including humans, plants, natural scenes, news, and architectures. The detailed specifications are listed in Table 1 and screenshots are shown in Fig. 3. Spatial information (SI) and temporal information (TI) [23] that roughly reflect the complexity of video content are also given in Table 1. Apparently, the video sequences are of diverse spatio-temporal complexity and widely span the SI-TI space. An 8-second video segment [3] is extracted from each source video, which is further partitioned into two non-overlapping 4-second segments, referred to as short segments (SS). We encode SS into 7 representations with H.264 encoder according to the encoding ladder shown in Table 2, where three compression levels, spatial resolutions, and frame rates are employed. A small-scale internal subjective test is conducted to divide the 7 representations into 3 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, we concatenated two consecutive 4-second segments with different representations from the same content into an 8-second long segment. We denoted the concatenated long segments by LS. Table 3 lists the quality adaptation patterns, from which we observe a diversity of adaptation intensities and types. Furthermore, to better exploit the space of adaptations, three types of multi-dimensional adaptations (Q-S, Q-T, and S-T) are also presented in the database. As a result, a total of 168 SS and 588 LS are included in the database.

3.2 Subjective User Study

Our subjective experiments generally follow the absolute category rating (ACR) methodology, as suggested by the ITU-T recommendation P.910 [23]. Although single-stimulus continuous quality evaluation (SSCQE) [23] is designed for continuously tracking instantaneous video quality over time, it is not adopted for the following reasons. First, human subjects are unlikely to evaluate video quality on a per frame basis in practice, discounting the instantaneous quality variations between frames within a scene. Second, in our database, the same coding configuration and parameters are applied to the full duration of each scene, which is roughly constant in

Table 3: 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; and 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	–

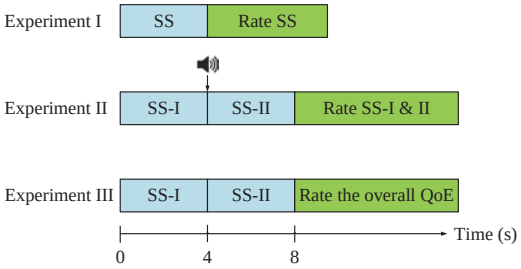


Figure 4: Experiment procedures. SS: short segment includes both SS-I and SS-II corresponding to the first 4-second and last 4-second video, respectively.

terms of content and complexity. As a result, a single score is sufficient to summarize its quality. Third, in SSCQE, there is time delay between the recorded instantaneous quality and the video content, and such delay varies between subjects and is also a function of the slider's "stiffness". This is an unresolved issue of the general SSCQE methodology, but can be avoided when a single score is acquired [15].

We carry out three subjective experiments as illustrated in Fig. 4. Subjects are invited to rate the quality of SS in Experiment I. The subjective rating of each SS is defined as the *intrinsic quality*. We perform Experiment II on LS, wherein subjects give two opinions to the first and second 4-second 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. In Experiment III, subjects are requested to watch the LS but to provide a single score to reflect their overall QoE. In order to remove any memory effects, we randomly shuffle content and adaptation patterns while ensuring that the same content and adaptation patterns are not consecutively displayed to a subject. A training session is performed before each experiment to familiarize subjects with typical distortion types and levels, so as to minimize the learning effect. 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 0-10 numerical quality scale suggested in the ITU-T recommendation P.910 [23].

The subjective testing is setup in a normal indoor home setting with an ordinary illumination level, with no reflecting ceiling walls and floors. All videos are displayed at their actual pixel resolution on an LCD monitor at a resolution of 1920 × 1080 pixels with Truecolor (32bit) at 60Hz. The monitor is calibrated in accordance with the ITU-T BT.500 recommendations [22]. 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. Visual acuity and color vision are confirmed from each subject before the test. Given the time constraints, each subject is assigned 5 out of the 12 contents in a circular fashion in Experiment III. To be specific, if subject i was assigned contents j to (j+4), then subject i+1 would watch contents (j+1) to (j+5). All 49 adaptation patterns for these contents are displayed to the subject only once.

4 DATA ANALYSIS

4.1 Data Processing

The subject rejection procedure in [22] is used and one subject is rejected from the experiment, resulting in 35 valid subjects. We average the scores from valid subjects to obtain the mean opinion score (MOS) for each video. We evaluate the performance of individual subjects by calculating the correlation between individual subject ratings and the MOSs for each video content, and then averaging correlations across content. The Spearman's rank-order correlation coefficient (SRCC) is employed as the comparison criterion, which ranges from 0 to 1 with a higher value indicating better performance. The mean and standard deviation of the results from Experiment I are shown in Fig. 5. The average performance across all subjects is also given in the rightmost column. Considerable agreement is observed among different subjects on the perceived quality of the test video sequences. Similar observations can be drawn from Experiment II and III, which are not reported here due to the page limit.

4.2 Experiment I

Fig. 6 plots the MOSs of SS with respect to the bitrate. Thanks to the initial subjective test before determining the

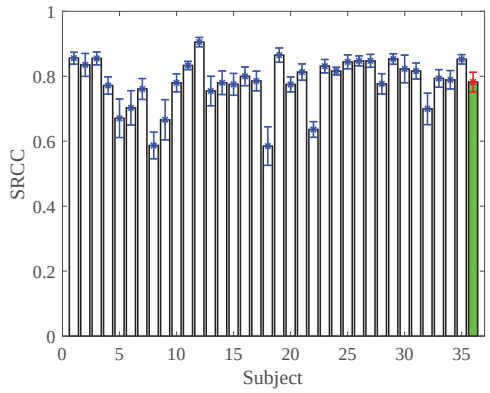


Figure 5: SRCC between the individual subject rating and the MOS. Rightmost column: performance of an average subject.

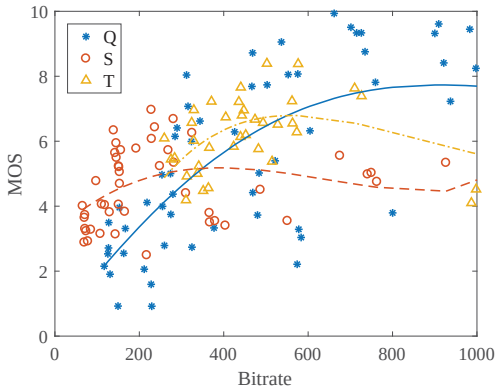


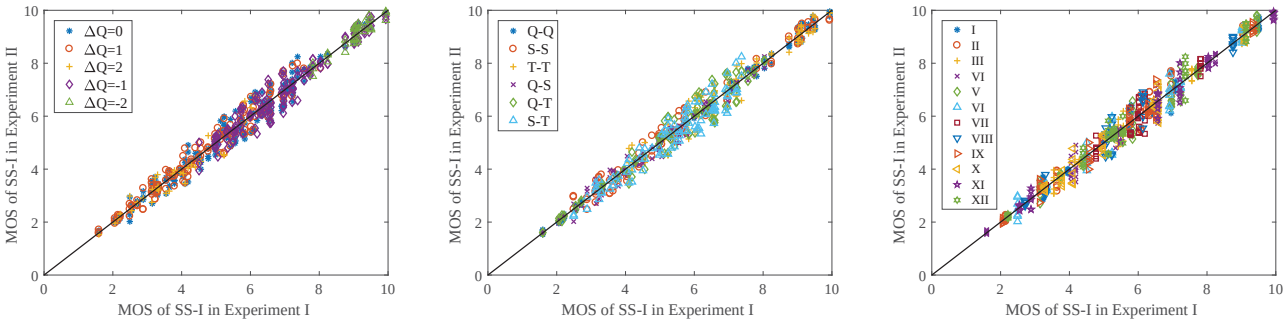
Figure 6: Rate-quality scatter plot. Q: H.264 compressed videos; S: spatially downsampled videos; T: temporally downsampled videos.

QP parameters used to create the compressed videos, the resulting MOS scatter in a wide range of the available scales, which allows us to study different cases of quality adaptation. We have several observations from the scatter plot. First, the MOS distribution of three distortion types has significant implications on the optimal encoding strategy of streaming videos. Specifically, encoding a low resolution video generally results in a better perceptual quality than encoding a high resolution video with the same resource at very low bitrates (*e.g.*, around 200kbps). Furthermore, frame rate reduction is not a wise choice for video compression. This may be because the temporal discontinuity caused by frame rate reduction increases the difficulty of motion prediction, which in turn increases the bitrate. As it comes to the medium bitrate range (*e.g.*, from 300kbps to 600kbps), the efficiency of frame rate reduction is higher than spatial resolution reduction, while pure H.264 compression achieves the best performance on average. When there is sufficient bandwidth resources (*e.g.* higher than 700kbps), light H.264 compression is preferred while neither spatial nor temporal resolution reduction is appreciated by subjects. Second, the optimal encoding strategy is largely content-dependent. For example, temporal resolution reduction does not affect the perceptual quality of the sequences 3dPrinter, Sunrise, and WildAnimal much, because those video sequences contain significantly less motion than others. On the other hand, the video sequences Chicago and MtTakao have more complex texture details, therefore more strongly affected by the loss of spatial resolution. Somewhat surprisingly, subjects prefer spatially downsampled videos than pure compressed videos with low spatial complexity such as FightForGlory and SplitTrailer, because compression often introduces unnatural stripe-shaped artifacts in smooth regions, while spatially down-sampling does not suffer from such problem. This suggests that the resource allocation scheme in H.264 encoder is not optimal in terms of perceptual quality, and a psychovisual rate-distortion enhanced encoder may achieve better performance. Given the observed content diversity, a one-size-fits-all encoding scheme obviously cannot provide the best video quality for a given title. Therefore, it is preferable to devise per-title encoding ladders for better resource allocation. Last, while Chicago and News have very similar temporal complexity according to the TI metric [23], the effect of frame rate reduction in the encoding process are drastically different, suggesting that a more sophisticated content-differentiating feature analysis is needed.

4.3 Experiment II

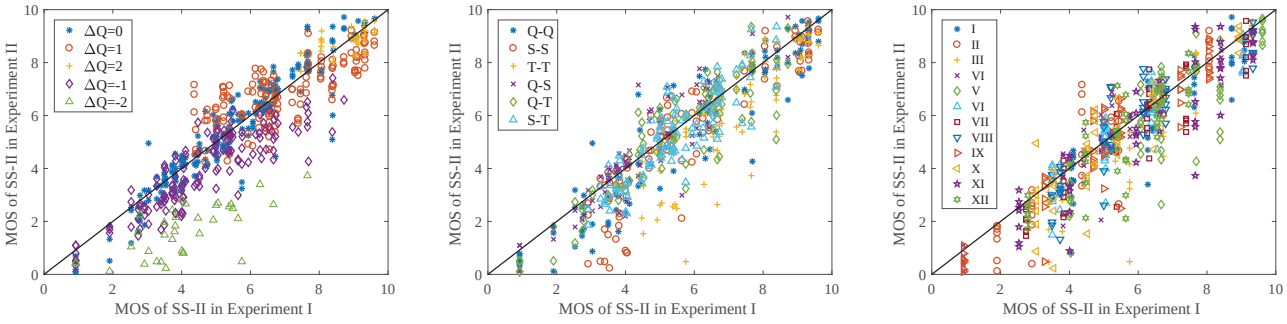
The *intrinsic quality* of SS are compared to the retrospective ratings of SS in Experiment II to investigate the influence of quality adaptations. As illustrated in Fig. 7 and Fig. 8, quality adaptations have substantially different impacts on the perceptual quality of video segments before and after switching. The MOSs on 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 deviation from *intrinsic quality*, and summarize the observations as follows.

Intensity effect: The intensity of quality change is the dominant factor of the perceptual quality deviation of SS-II. Fig. 8(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 negative quality adaptation. One explanation may be that there is a higher viewer expectation when viewers are exposed to high video quality 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 [9–12, 29]. On the other hand, we do not observe a consistent penalty or reward across all ranges for constant or positive quality adaptation scenarios.



(a) Intensity (b) Type (c) Content

Figure 7: MOS of the SS-I in Experiment I vs. MOS of the SS-I in Experiment II.



(a) Intensity (b) Type (c) Content

Figure 8: MOS of the SS-II in Experiment I vs. MOS of the SS-II in Experiment II.

Type effect: The type of adaptation, given in Table 3, is another major influential factor of QoE. Significant differences between subjective ratings given to different types of adaptation can be found in Fig. 8(b). In particular, temporal resolution adaptation is rated as the least favorable approach in the quality adaptation, even in the case of positive adaptation. Compression level and spatial resolution adaptations, on the other hand, do not introduce extra penalty in general, whereas subjects penalize sudden occurrence of blurring artifact when the quality of SS-I is high. It is also noteworthy that the multi-dimensional adaptation types Q-T and S-T also introduce penalty on the perceptual quality of SS-II, especially when the *intrinsic quality* ranges from medium to high, while the Q-S adaptation does not have such effect.

Level effect: It is suggested that the amount of reward or penalty that subjects give to SS-II is not only affected by the intensity effect and type effect, but also the *intrinsic quality* level where the adaptation occurs. The vertical distance from the green points to the diagonal line in Fig. 8(a) increases along the horizontal axis, suggesting a quality degradation occurred in the 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 [2]. However, the amount of reward is relatively small comparing to the penalty introduced by negative quality adaptation, 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 consistent results in quality adaptations. The reason behind is not fully understood but is worth deep investigation.

Content effect: Content seems to play a minor role in quality adaptations. However, we observe that the contents without scene change such as Chicago and StreetDance are more heavily degraded by quality adaptations than the video sequences consisting 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.

We further perform ANalysis Of VAriance (ANOVA) test on the MOSs of SS-II from Experiment II to understand the statistical significance of the influencing factors, where the p-value is set to 0.05. The results suggest that intensity of adaptation, type of adaptation, *intrinsic quality* level,

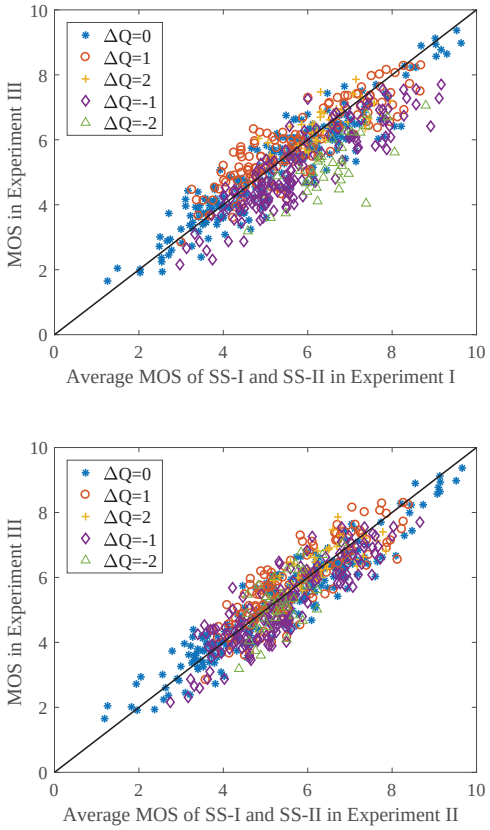


Figure 9: Performance of average pooling in Experiment I and II.

content, the interactions between intensity and type of adaptation, and the interactions between *intrinsic quality* level and intensity of adaptation are statistically significant to the MOS discrepancy between Experiments I and II.

4.4 Experiment III

To understand the strategy that subjects employed to integrate segment-level perceptual video quality into an overall QoE score, we evaluate 7 temporal pooling strategies using both the *intrinsic quality* and *post-hoc* quality obtained in Experiment I and II, respectively [15]. These include Mean, Min, and Max MOS of all segments, the MOS of SS-I and SS-II, weighted average MOS with increasing weights (W+), where $w = [\frac{1}{3}, \frac{2}{3}]$ for long video clips, and decreasing weights (W-), where $w = [\frac{2}{3}, \frac{1}{3}]$. SRCC and Pearson linear correlation coefficient (PLCC) between the predicted and actual QoE are then calculated to provide quantitative evaluation, shown in Table 4 and Table 5, respectively. It is interesting to note that the average pooling of *post-hoc* segment-level scores exhibit the highest correlation to the overall QoE, even outperforming the increasing weights pooling strategy that is designed to account for the *recency effect*. To ascertain that the improvement of the *post-hoc* average model is statistically significant, we performed F-test on the prediction residual as suggested in [11]. We observe the *post-hoc* average model is significantly better than the W+ model with a 95% confidence level. This suggests that the impact of the *recency effect* is secondary in adaptive streaming. Furthermore, Fig. 9 compares the scatter plots of the MOS versus the average intrinsic segment-level quality and the average *post-hoc* segment-level MOS, respectively. The average *intrinsic quality* tends to overestimate the QoE of LS with negative quality adaptation while the average *post-hoc* MOS achieves better performance in predicting the overall QoE. This observation suggests a promising approach in developing objective QoE models: instead of applying sophisticated temporal pooling strategies, we should first predict the *post-hoc* segment-level video quality, which is affected by the four influencing factors of quality adaptation. Average pooling on the *post-hoc* quality scores is then sufficient to predict the overall QoE.

5 PERFORMANCE OF OBJECTIVE VQA MODELS

We test 5 objective VQA models including PSNR, SSIM [26], MS-SSIM [27], SSIMplus [16], and VQM [14] along with 7 temporal pooling strategies as described in Section 4.4. To be specific, for each objective VQA algorithm, we compute the objective VQA scores for each SS by averaging frame-level scores, resulting in 168 predicted quality scores. We then apply the temporal pooling schemes on the segment-level video quality scores in each LS. Since none of the VQA algorithms supports cross-resolution and cross-frame-rate video quality evaluation except for SSIMplus, we up-sampled all representations to 1920×1080 and 30 fps, and then apply the VQA on the up-sampled videos because it is the size of display in the subjective experiment. Table 4 and Table 5 summarize the evaluation results on LS, which are somewhat disappointing because state-of-the-art VQA models and temporal pooling schemes do not seem to provide adequate predictions of perceived time-varying video quality.

The test results also provide some useful insights regarding the general approaches used in VQA models. First, from the significant improvement of MS-SSIM and SSIMplus upon SSIM, we may conclude that the multi-scale approach performs better against variations in resolution. 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 covered in traditional VQA models. Third, although average pooling of intrinsic segment-level quality results in suboptimal performance, none of the existing pooling strategies outperforms average pooling consistently because subjects employ asymmetric strategies in updating their opinions. The approach suggested in Section 4.4 has the potential to greatly improve the performance of VQA in predicting time-varying video quality.

Table 4: SRCC comparison between actual MOS and predicted MOS using different base quality measures (segment-level MOS, *post-hoc* MOS, segment-level PSNR, segment-level SSIM, segment-level MS-SSIM, segment-level SSIMplus, and segment-level VQM) and different pooling strategies (mean, min, max, median, SS-I, SS-II, W+, and W-). Segment-level objective VQA is computed as frame average.

Base measure	PSNR	SSIM [26]	MS-SSIM [27]	SSIMplus [16]	VQM [14]	Segment-level MOS	*Post-hoc* MOS
Mean	0.30	0.23	0.42	0.70	0.59	0.85	**0.90**
Min	0.21	0.18	0.36	0.58	0.54	0.75	0.73
Max	0.35	0.31	0.51	0.60	0.51	0.71	0.68
SS-I	0.16	0.14	0.27	0.32	0.32	0.42	0.42
SS-II	**0.37**	**0.34**	**0.55**	0.65	**0.64**	0.78	0.68
W+	0.33	0.26	0.45	**0.75**	0.62	**0.87**	0.86
W-	0.25	0.20	0.38	0.59	0.51	0.72	0.76

Table 5: PLCC comparison between actual MOS and predicted MOS after a non-linear mapping using different base quality measures (segment-level MOS, *post-hoc* MOS, segment-level PSNR, segment-level SSIM, segment-level MS-SSIM, segment-level SSIMplus, and segment-level VQM) and different pooling strategies (mean, min, max, median, SS-I, SS-II, W+, and W-). Segment-level objective VQA is computed as frame average.

Base measure	PSNR	SSIM [26]	MS-SSIM [27]	SSIMplus [16]	VQM [14]	Segment-level MOS	Post-hoc MOS
Mean	0.32	0.27	0.49	0.73	0.59	0.87	**0.90**
Min	0.26	0.19	0.47	0.64	0.53	0.79	0.77
Max	0.36	**0.33**	0.50	0.62	0.52	0.72	0.70
SS-I	0.18	0.16	0.29	0.36	0.33	0.45	0.45
SS-II	**0.39**	**0.33**	**0.56**	0.68	**0.62**	0.79	0.69
W+	0.35	0.31	0.53	**0.77**	**0.62**	**0.88**	0.87
W-	0.28	0.22	0.44	0.62	0.52	0.75	0.79

6 CONCLUSIONS

To study the visual QoE in adaptive video streaming, we introduce a new database involving 168 short and 588 long video clips with variations in compression level, spatial resolution, and frame-rate. We design a series of subjective experiments to exploit the multi-dimensional adaptation space. The database, together with the subjective data, will be made available to the public to facilitate future QoE research. Our path-analytic results indicate that the perceptual quality of the former video segment has a direct effect on the subsequent video segment, which in turn influences the overall QoE. The perceptual quality deviation introduced by quality adaptation is a function of the intensity of adaptation, type of adaptation, intrinsic video quality, content, and the interactions between them. We find that positive quality adaptation exhibits a Weber's law effect, but the effect is less significant for negative adaptations. The opinion updating function has the potential to greatly improve the performance of objective VQA models, which by themselves do not deliver strong prediction results.

Many challenging problems remain to be solved. First, due to the limited duration 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, very little work has been dedicated to explain the underlying mechanism why quality variation affects QoE. A theoretical framework that integrates the useful findings from the current study may guide the development of subjective and objective QoE models for streaming videos. Third, optimization of the existing video streaming frameworks based on the current findings is another challenging problem that desires further investigations.

REFERENCES

[1] DASH Industry Forum. 2013. For Promotion of MPEG-DASH. (2013). Retrieved March 30, 2017 from http://dashif.org.
[2] G.T. Fechner. 2012. *Elements of psychophysics*. Vol. 2. Breitkopf und Hrtel.
[3] P. Fröhlich, S. Egger, R. Schatz, M. Mühlegger, K. Masuch, and B. Gardlo. 2012. QoE in 10 seconds: Are short video clip lengths sufficient for Quality of Experience assessment?. In *Proc. IEEE Int. Conf. on Quality of Multimedia Experience*. 242–247.
[4] M.N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnstrom, and A. Raake. 2014. Quality of Experience and HTTP adaptive streaming: A review of subjective studies. In *Proc. IEEE Int. Conf. on Quality of Multimedia Experience*. 141–146.
[5] M. Grafl and C. Timmerer. 2013. Representation switch smoothing for adaptive HTTP streaming. In *IEEE Int. Workshop on Perceptual Qualtiy of Systems*. 178–183.
[6] D.S. Hands and S.E. Avons. 2001. Recency and duration neglect in subjective assessment of television picture quality. *Applied*

Cognitive Psychology 15, 6 (Nov. 2001), 639–657.

[7] B. Lewcio, B. Belmudez, A. Mehmood, M. Wältermann, and S. Möller. 2011. Video quality in next generation mobile networks-perception of time-varying transmission. In *IEEE Int. Workshop Technical Committee on Comm. Quality and Reliability*. 1–6.

[8] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao. 2015. Deriving and validating user experience model for dash video streaming. *IEEE Trans. Broadcasting* 61, 4 (Dec. 2015), 651–665.

[9] Y. Liu, Y. Shen, Y. Mao, J. Liu, Q. Lin, and D. Yang. 2013. A study on Quality of Experience for adaptive streaming service. In *Proc. IEEE Int. Conf. Comm. Workshop*. 682–686.

[10] R.K.P. Mok, E.W.W. Chan, and R.K.C. Chang. 2011. Measuring the Quality of Experience of HTTP video streaming. In *Proc. IFIP/IEEE Int. Sym. Integrated Network Management*. 485–492.

[11] A.K. Moorthy, L.K. Choi, A.C. Bovik, and G. De Veciana. 2012. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing* 6, 6 (Oct. 2012), 652–671.

[12] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen. 2011. Flicker effects in adaptive video streaming to handheld devices. In *Proc. ACM Int. Conf. Multimedia*. 463–472.

[13] Y. Ou, Y. Xue, and Y. Wang. 2014. Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions. *IEEE Trans. Image Processing* 23, 6 (June 2014), 2473–2486.

[14] M.H. Pinson and S. Wolf. 2004. A new standardized method for objectively measuring video quality. *IEEE Trans. Broadcasting* 50, 3 (Sept. 2004), 312–322.

[15] A. Rehman and Z. Wang. 2013. Perceptual experience of time-varying video quality. In *Proc. IEEE Int. Conf. on Quality of Multimedia Experience*. 218–223.

[16] A. Rehman, K. Zeng, and Z. Wang. 2015. Display device-adapted video Quality-of-Experience assessment. In *Proc. SPIE*. 939406.1–939406.11.

[17] D.C. Robinson, Y. Jutras, and V. Craciun. 2012. Subjective video quality assessment of HTTP adaptive streaming technologies. *Bell Labs Technical Journal* 16, 4 (Mar. 2012), 5–23.

[18] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hobfeld, and P. Tran-Gia. 2014. A survey on Quality of Experience of HTTP adaptive streaming. *IEEE Communications Surveys & Tutorials* 17, 1 (Sept. 2014), 469–492.

[19] T. Stockhammer. 2011. Dynamic adaptive streaming over HTTP: Standards and design principles. In *Proc. ACM Conf. on Multimedia Systems*. 133–144.

[20] J. Talens-Noguera, W. Zhang, and H. Liu. 2015. Studying human behavioural responses to time-varying distortions for video quality assessment. In *Proc. IEEE Int. Conf. Image Proc.* 651–655.

[21] S. Tavakoli, K. Brunnström, K. Wang, B. Andrén, M. Shahid, and N. Garcia. 2014. Subjective quality assessment of an adaptive video streaming model. In *IS&T/SPIE Electronic Imaging*. 9016.1–9016.13.

[22] International Telecommuniations Union. 1993. ITU-R BT.500-12. In *Recommendation: Methodology for the subjective assessment of the quality of television pictures*.

[23] International Telecommuniations Union. 1999. ITU-R BT.910. In *Subjective video quality assessment methods for multimedia applications*.

[24] B.J. Villa, K. De Moor, P.E. Heegaard, and A. Instefjord. 2013. Investigating Quality of Experience in the context of adaptive video streaming: Findings from an experimental user study. In *Akademika forlag Stavanger*. 122–133.

[25] Z. Wang and A.C. Bovik. 2006. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing* 2, 1 (Dec. 2006), 1–156.

[26] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing* 13, 4 (Apr. 2004), 600–612.

[27] Z. Wang, E.P. Simoncelli, and A.C. Bovik. 2003. Multiscale structural similarity for image quality assessment. In *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*. 1398–1402.

[28] M. Zink, O. Künzel, J. Schmitt, and R. Steinmetz. 2003. Subjective impression of variations in layer encoded videos. In *International Workshop on Quality of Service*. Springer, 137–154.

[29] M. Zink, J. Schmitt, and R. Steinmetz. 2005. Layer-encoded video in scalable adaptive streaming. *IEEE Trans. Multimedia* 7, 1 (Feb. 2005), 75–84.