Study of Temporal Effects on Subjective Video Quality of Experience

Christos George Bampis, *Student Member, IEEE*, Zhi Li, *Member, IEEE*, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*— HTTP adaptive streaming is being increasingly deployed by network content providers, such as Netflix and YouTube. By dividing video content into data chunks encoded at different bitrates, a client is able to request the appropriate bitrate for the segment to be played next based on the estimated network conditions. However, this can introduce a number of impairments, including compression artifacts and rebuffering events, which can severely impact an end-user's quality of experience (QoE). We have recently created a new video quality database, which simulates a typical video streaming application, using long video sequences and interesting Netflix content. Going beyond previous efforts, the new database contains highly diverse and contemporary content, and it includes the subjective opinions of a sizable number of human subjects regarding the effects on QoE of both rebuffering and compression distortions. We observed that rebuffering is always obvious and unpleasant to subjects, while bitrate changes may be less obvious due to content-related dependencies. Transient bitrate drops were preferable over rebuffering only on low complexity video content, while consistently low bitrates were poorly tolerated. We evaluated different objective video quality assessment algorithms on our database and found that objective video quality models are unreliable for QoE prediction on videos suffering from both rebuffering events and bitrate changes. This implies the need for more general QoE models that take into account objective quality models, rebuffering-aware information, and memory. The publicly available video content as well as metadata for all of the videos in the new database can be found at http://live.ece.utexas.edu/research/LIVE_NFLXStudy/nflx_index.html.**

***Index Terms*— Subjective quality of experience, video quality assessment, video streaming.**

I. Introduction

GLOBAL mobile data traffic grew 74% and mobile video traffic accounted for 55 percent of total mobile data traffic in 2015 [1]. According to the Cisco Visual Networking Index and global mobile data traffic forecast, mobile data traffic will grow 8-fold from 2015 to 2020, which constitutes a compound annual growth rate of 53%. Adding to the delivery over fixed networks, this large and growing volume of mobile video data, video streaming providers such as Netflix, Youtube and Hulu are processing, storing and delivering vast amounts of video data on a daily basis. Given the exploding use of mobile video devices and the tremendous network bandwidth demands of streaming users, the biggest challenge in video content delivery is to create better network-aware strategies to improve end-users' quality of experience (QoE). In this direction, HTTP Adaptive Streaming (HAS) is being used by content providers as a way of dealing with network fluctuations.

The core idea of HAS is that every video content is divided into chunks that are each encoded using a different set of parameters. The client side then decides which bitrate to use for the chunk to be played next, given a set of critical parameters, such as the estimated network conditions over the next few seconds, buffer size [2], etc. Given that HAS uses TCP as the transfer protocol, video impairments such as missing frames or portions of frames, due to error-prone wireless networks and packet loss are eliminated. However, when the available bandwidth is low and the client buffer is empty, start-up delays and rebuffering events may occur. This approach leads to impairments including frozen video frames - the result of a rebuffering event - and/or highly visible compression artifacts. Given that the end goal of every content provider is to maximize the end-user's QoE while mediating parameters to accommodate network changes and changing bandwidth, subjective modelling of streaming video QoE becomes an important objective.

Subjective testing is an established way of measuring QoE under different scenarios and settings. Many successful studies have been developed using short video sequences of 10-15 seconds (or even less) as in [3]–[6]. However, these studies do not reflect typical video streaming situations, where subjects view videos that could be minutes long. Hence, it is not possible to analyze long-term memory effects as they relate to critical factors affecting subjective QoE such as the recency effect [7].

Longer video sequences were considered in [8], where video delivery over HAS was simulated on tablet devices. The authors studied combinations of bitrate changes and rebuffering events, but their analysis was limited to 6 sequences, 3 playout scenarios and 26 subjects. Longer video sequences were also used in [9], using video contents ranging from 30 to 60 sec. The authors studied the effect of rebuffering

Manuscript received January 19, 2017; revised April 28, 2017; accepted July 7, 2017. Date of publication July 20, 2017; date of current version August 21, 2017. This research was supported in part by a grant from Netflix Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo. *(Corresponding author: Christos George Bampis.)*

C. G. Bampis and A. C. Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78705 USA (e-mail: bampis@utexas.edu; bovik@ece.utexas.edu).

Z. Li, A. K. Moorthy, I. Katsavounidis, and A. Aaron are with Netflix Inc., Scotts Valley, CA 95032 USA.

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.2017.2729891

events as a function of location and density in a video sequence. However, temporal ratings were not collected, hence their analysis was based only on a final summary rating (retrospective score). As we will show later, retrospective ratings tend to be affected by recency biases. The study of temporal pooling techniques in [10] also included longer video sequences, and concluded that current temporal pooling strategies are mostly effective on short videos. On the long videos they used, simple mean pooling was found to be superior to all other methods. However, they only used two video contents in their analysis.

In all previous studies, playout patterns were chosen without considering the role of bandwidth usage. Generally, subject rejection strategies have been based only on retrospective scores or conducted on a per frame basis. We argue that such methodologies are inappropriate when gathering temporal scores, particularly when studying the complex temporal effects that affect subjective QoE.

To sum up, previous efforts suffer from at least one of the following:

1) a small number of contents, playout patterns or number of subjects
2) a lack of practical network or buffer constraints on the subjective test design
3) use of short video sequences that do not capture long term temporal effects
4) not including both temporal and retrospective QoE scores
5) not deploying temporal subject rejection methods

Here, we describe a set of experiments that we conducted to gather data that will help us develop tools to create perceptually optimized network allocation protocols. We conducted experiments to measure subjective QoE in a typical mobile video streaming setting, where the human subjects were exposed to diverse real-world content, realistic network conditions and client-based strategies, while viewing video sequences of durations of at least one minute, displayed on a small mobile screen at low bitrates.

The outcome of these experiments is the new LIVE-Netflix mobile VQA database, which consists of 112 distorted videos evaluated by over 55 human subjects on a mobile device. The distorted videos were generated from 14 video contents of spatial resolution 1080p at 24, 25 and 30 fps by imposing a set of 8 different playout patterns including: dynamically changing H.264 compression rates, rebuffering events and mixtures of both. While more recent compression standards such as H.265/HEVC and VP9 are currently being developed, H.264 is currently the most widely used format. Further, while H.265 achieves higher efficiency than H.264 does, it is not conceptually different from H.264: it uses the same motion-compensated/transform/lossless entropy coding hybrid model and essentially the same coding tools. Therefore, coding artifacts are perceptually similar among these two codecs; we thus expect the results of this study to apply to H.265-based streaming, with appropriately lowered encoding bitrates.

The database contains 11 different types of content provided by Netflix (drama, action, comedy, anime etc.) and 3 publicly available video contents from the Consumer Digital Video Library (CDVL) [11]. To provide a more realistic viewing experience, the audio track was included and played without distortion when the subjects viewed each sequence. Figure 1 shows an example of the type of impairments introduced on the videos in the LIVE-Netflix Dataset.

Fig. 1. Network impairment simulation using H.264 compression (left) and rebuffering events (right). The red box indicates a compression artifact.

Given the lack of available subjective datasets driven by practical network constraints or streaming client strategies, our goal was to design a dataset of significant practical value. Hence, we designed the LIVE-Netflix dataset based on playout scenarios that are common when streaming under practical bandwidth constraints and buffer size limitations. We also gathered both continuous and retrospective QoE scores towards achieving a more complete understanding of how humans combine different aspects of temporal perception into a single, overall impression of QoE. We believe that this work offers the possibility to bring human behavior modeling in this context closer to traditional video quality assessment (VQA) research. To both demonstrate the value of the database, as well to provide an engineering comparison of practical worth, we evaluated various state-of-the-art VQA algorithms and temporal pooling strategies on the new database. We also extensively studied temporal effects on subject QoE by analyzing the collective and per video impairment behavior of the subjects.

Our analysis led us to draw various observations. First, we observed that rebuffering severely affected subject QoE regardless of the content. Therefore, subjects tended to prefer transient bitrate drops over rebuffering on low complexity contents, even when the selected bitrate was low. However, a constant low bitrate - to avoid rebuffering - was not tolerated by subjects. Finally, the gathered subjective data strongly manifested known QoE phenomena such as the recency effect (more recent video segments have a disproportionate effect on perceived visual quality) and the non-linearity of human responses, but it also challenges the use of retrospective scores or global subject rejection methodologies for QoE assessment on long videos.

The rest of this paper is organized as follows. Section II describes the dataset design, the encoding pipeline and the source contents used. Section III presents the subjective testing methodology, and Section IV discusses the processing of subjective scores and the proposed subject rejection method. Sections V and VI analyze the collected retrospective and continuous QoE scores, while Section VII explores the cognitive aspects of subjective QoE in light of the collected human data. Section VIII analyzes the performance of various VQA algorithms and Section IX gives conclusions.

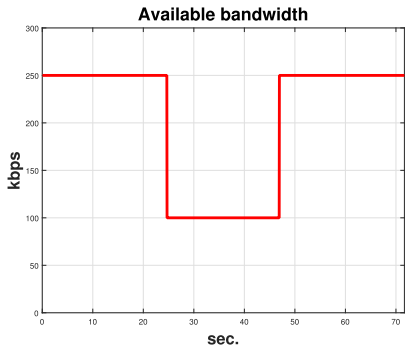


Fig. 2. Available bandwidth model used in the LIVE-Netflix dataset. All of the test sequences were designed to consume the same amount of network resources (bandwidth).

II. Subjective Assessment of Mobile Video Quality

A. Network Assumptions and Buffer Limitations

When designing resource allocation strategies, content providers seek to answer the question: given a fixed amount of network resources, which strategy delivers the highest possible QoE? We consider here the tradeoffs that occur on end-users' QoE when mediating between rebuffering events and bitrate reduction under a mobile low bitrate regime. To do so, we designed a set of realistic playout patterns, assuming the same network resources and same buffer limitations. To simulate realistic network conditions, we used a channel with time-varying capacity, shown in Fig. 2. The available bandwidth starts at 250 kbps, followed by a temporary bandwidth drop to 100 kbps of duration $d = 22.2167$ seconds until the bandwidth recovers to its previous 250 kbps value. This simple example of a bandwidth drop can be used as a building block to simulate models of more complex network conditions. Using this available bandwidth model, we derived eight test patterns based on the premise that the average playout rate of the client side cannot exceed that of the average bandwidth. The only exception to this rule is when the client uses some of the available buffer. Next, we discuss the buffer usage aspects of the designed patterns.

To ensure the practical worth of the derived sequences, it is necessary to take into account the available buffer size. As shown in [2], a buffer-based strategy can be a simple and useful way to reduce the number of rebuffering events and bitrate switches that occur. Clearly, there are three possibilities:
1. The (instantaneous) playout rate is smaller than the (instantaneous) available bandwidth; the buffer is being filled with more data.
2. The playout rate is larger than the available bandwidth; the buffer is being emptied.
3. The playout rate is equal to the available bandwidth; the buffer state does not change over time.

Given our network assumption, we also considered a specific initial buffer state for streaming, where the buffer of size B_0 was filled with video chunks encoded at 250 kbps. We further assumed two possible initial buffer states: $B_0 = 1333$ kbits or $B_0 = 0$ kbit. The former scenario corresponds to "steady state" streaming where the initial buffer is filled, while the latter assumes that there is no initial buffer available. All patterns were designed so that the buffer is emptied at the end of the bandwidth drop shown in Fig. 2.

B. Playout Patterns

Based on the aforementioned network scenario and possible values for B_0, we simulated the following client approaches (see also Fig. 3 for an overview):

0) A constant encoding bitrate of 500 kbps. This playout pattern assumes an impairment-free network condition where the bandwidth is sufficient to allow such a playout rate by the client. In this case, the buffer is not used at all. This pattern is the only one that does not satisfy the bandwidth and buffer constraints. Although we included this pattern among the viewed playout patterns, it did not serve as a "hidden reference" [3].
1) One video chunk encoded at 250 kbps followed by an 8 sec. stall, followed by another 250 kbps chunk (see Fig. 4). The client drains the buffer completely before the rebuffering event occurs. Before the available bandwidth recovers, the client decides to resume playback after the 8 second rebuffer. By the end of the pattern, the buffer is emptied.
2) A single video chunk of $R_2 = 160$ kbps. The client side is very conservative throughout the video playback by always picking a playout rate of R_2, so that there is no rebuffering and the available buffer is depleted.
3) One video chunk encoded at 195 kbps, followed by a 4 sec. stall, followed by another 195 kbps chunk. Here, the client strategy is to reduce the rebuffering duration by half (4 sec.), by using a lower encoding bitrate. As before, during the rebuffering event, the client has a zero playout rate but an encoding bitrate of 100 kbps (equal to the available bandwidth) which allows the buffer level to partially recover and then be used to stream at 195 kbps before bandwidth recovers (see also Fig. 4).
4) One video chunk encoded at 250 kbps followed by a 66 kbps chunk, followed by another 250 kbps chunk. This playout pattern is an alternative to pattern #1, where the client tries to avoid any rebuffering events by switching to a lower playout rate (66 kbps) than the available bandwidth (100 kbps) during the bandwidth drop.

By removing the assumption on the availability of the buffer on the client side ($B_0 = 0$), a second set of playout patterns can also be simulated. This set of patterns is likely to deliver lower QoE scores to subjects since more severe impairments have to be introduced to deal with the bandwidth drop.

5) One video chunk at 250 kbps, followed by a 6.66 sec. rebuffering event, followed by a chunk at 250 kbps, followed by another 6.66 sec. rebuffering event, followed by the last 250 kbps chunk. In pattern #5, the unavailability of the buffer leads to rebuffering. By filling some of the buffer, the client is able to play out for a small interval of time at 250 kbps until the buffer is depleted.

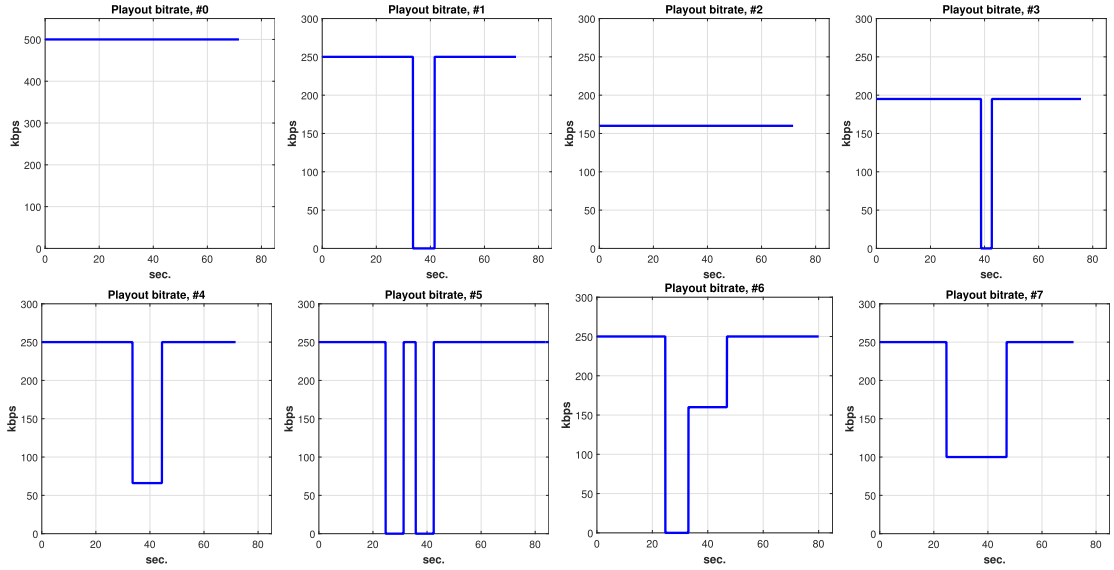


Fig. 3. Playout patterns used in the subjective study. First row: patterns #0 until #3, second row: patterns #4 until #7. The horizontal axis corresponds to frame indices while the vertical corresponds to the instantaneous playout bitrate in kbps.

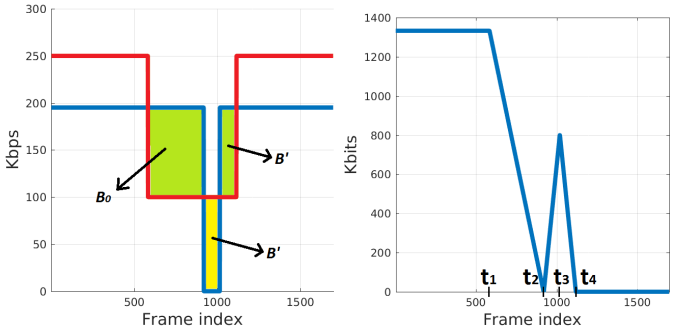


Fig. 4. Left: Blue denotes the playout pattern #3 while red denotes the available bandwidth. The green areas correspond to buffer consumption while the yellow area indicates the buffer build-up. B_0 corresponds to the available buffer at the beginning of the bandwidth drop, while B' corresponds to the amount of buffered data being filled, then consumed by the client. Right: Available buffer level over time for playout pattern #3, [t_1 t_2]: buffer drainage, [t_2 t_3]: buffer build-up, [t_3 t_4]: buffer drainage.

This leads to the rebuffering event, which is followed by a recovery at 250 kbps playout over a small time interval until the bandwidth also recovers.

6) One video chunk at 250 kbps, followed by a 8.33 sec. rebuffering event, followed by a chunk at 160 kbps, then a final video chunk at 250 kbps. Here, the client seeks to avoid a second rebuffering event by a gradual bitrate recovery.
7) One video chunk at 250 kbps is followed by a chunk at 100 kbps and then another chunk encoded at 250 kbps. Here it is assumed that the client is immediately able to adjust to the network conditions by using a playout rate that is always equal to the available bandwidth/encoding bitrate. This pattern may be the least practical among all the considered playout patterns. However, it is of interest to be able to study the subjective data resulting from such an "ideal" client reaction.

In the Appendix, we give an example of how some of the previous parameters were specified. Note that the original video sequences were of different durations, and that the playout patterns (of a given content) may also be of different durations because of delays introduced by rebuffering events.

C. Encoding Pipeline

We developed an encoding pipeline that generates the different parts of the final video and appropriately concatenates them based on an encoding map that indicates the time intervals of every quality level, the location and the duration of each rebuffering event. First, the source video stream (in H.264 format) was decoded, yielding an uncompressed raw .yuv file. The encoding map was then used to split the .yuv file in a frame-accurate manner, yielding .yuv chunks. A two pass encoding step using FFMPEG [12] was then used to encode the .yuv files into .mp4 format. Meanwhile, the final frame of a video chunk immediately before a rebuffering event was used to generate a rebuffering video chunk. A customized "loading", or spinning wheel, icon was overlaid on that frame and appropriately animated to simulate the desired video rebuffering effect. For playback purposes, and in order to match the rendering device resolution, all YUV frames were first upsampled to 1920 × 1080. An MP4 file was then created by lightly compressing these frames at CRF [13] (constant-rate-factor) value of 10. A more detailed description of the encoding pipeline that we used can be found in the Appendix.

D. Source Contents

A set of 14 video test contents were used containing a wide variety of spatiotemporal characteristics. Of the 14 contents, 11 belong to the Netflix catalog of titles including action scenes, drama, adventure, anime and cartoons. The remaining 3 contents were obtained from the publicly available Consumer Digital Video Library (CDVL) [11]. A few frames from

Fig. 5. Some frames from the LIVE-Netflix dataset. From left to right: ElFuente and Chimera sequences from the dataset.

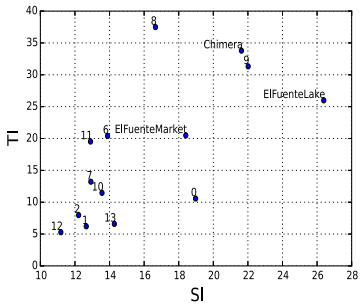


Fig. 6. Spatial Information (SI) plotted against Temporal Information (TI) for the 14 video test contents in the LIVE-Netflix dataset.

the video sequences are shown in Fig. 5. The test contents have a variety of frame rates and resolutions. For example, the ElFuente sequence has 4K resolution (4096×2160) and a frame rate of 60 fps, whereas most of the videos from the Netflix catalog have 1080p (1920×1080) resolution and frame rates of either 24, 25 or 30 fps. To deal with this difference, the ElFuente sequence was downscaled to 1080p and the frame rate was converted from 60 fps to 30 fps.

Measurements of spatial and temporal complexity give a rough idea of the content variety in a subjective database [14]. Let F_n denote the luminance channel of a video frame at time n and (i, j) the spatial coordinates of this frame. Next, consider the following simple Spatial Information (SI) and Temporal Information (TI) metrics [15]:

$$\text{SI} = \max_n \left\{ \text{std}_{i,j}[\text{Sobel}(F_n)] \right\}, \quad \text{TI} = \max_n \left\{ \text{std}_{i,j}[M_n(i, j)] \right\}$$

where $M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$, $\text{std}_{i,j}(.)$ denotes the standard deviation over all pixels (i, j) and $\max_n$ denotes the maximum over all frames. As shown in Fig. 6, the video content we use widely spans the SI-TI space [15].

III. Subjective Testing

A. Subjective Study Design

A single-stimulus continuous quality evaluation study [16] was conducted over a period of three weeks at The University of Texas at Austin's LIVE subjective testing lab. We collected subjective data from 56 subjects and a total of 4928 continuous scores together with the corresponding retrospective scores. Visual fatigue is an important consideration when designing subjective studies, so we split the study into three sessions, spaced by at least 24 hours to minimize subject fatigue [16]. Each session contained video content at most 35 minutes long, and the overall duration of each session was about 45 minutes.

Due to necessary limitations on the duration of a subjective study, video QoE studies invariably must limit the number of different contents that are shown. When using longer video sequences, this is even more challenging. Driven by a desire to deploy as diverse and large set of contents as possible, we employed the following strategy. Each subject was assigned 11 contents (of the 14) in a circular fashion e.g. if subject i as assigned contents 1 through 11, then subject $i + 1$ watched contents 2 through 12. This could result in a slightly different number of temporal and retrospective scores per content, but given the large number of subjects, we deemed this to be a statistically insignificant difference. All 8 playout patterns for these 11 contents were displayed to the subject only once. In order to remove any memory effects, we randomly shuffled the contents and the corresponding playout patterns while ensuring that the same content was not consecutively displayed to a subject in any session.

Android Studio was used to modify an earlier version of the human subject interface used in [4], which was made available to us by the authors. Using the previously described encoding pipeline, the generated .mp4 files were displayed on a Samsung S5 mobile device with a 1080p resolution and 5.1" screen size. This device had no problems playing the videos which were stored locally on an external SD card. The use of an external SD card did not introduce any latency when displaying the videos. The mobile device was not calibrated, but the brightness level was held constant at approximately 75% of maximum throughout the study. The sampling rate on the continuous scores was such that one score was measured per frame. Given the different frame rates of the input sequences, we parameterized the number of samples per video content depending on each video's frame rate.

B. Subjective Testing Walkthrough

Here we describe subjective testing procedure as it occurred during the first (training) session of each subject. Once seated, each subject was briefly instructed regarding the subjective testing process. They were asked to rate both their continuous and their overall QoE based on everything that they viewed on the screen. They were also asked not to make QoE judgments based on the level of interestingness of the video content or the audio quality. To remove any rating biases, the subjects were informed that there were no right or wrong answers in the experiment. No formal visual acuity test was performed, but the subjects verbally verified that they had normal or corrected-to-normal acuity. If a subject normally used corrective lenses when watching videos, they were asked to use them during the study.

Then, the subjects were introduced to the interface and the different video impairments they would be exposed to. Three different video contents, each with a different playout pattern were displayed as each subject became familiar with the testing interface. These contents were the same for all subjects but were not among the test contents used to gather the subjective data. After the first session, no training videos were shown, since subjects were assumed to be adequately familiar with the testing procedure and interface.

The video sequences were displayed one after the other and a continuous scale rating bar was displayed at the bottom of the mobile device screen. The ratings on the continuous (Likert)

Fig. 7. Subjective testing interfaces. Left: continuous QoE scoring; Right: retrospective scoring.

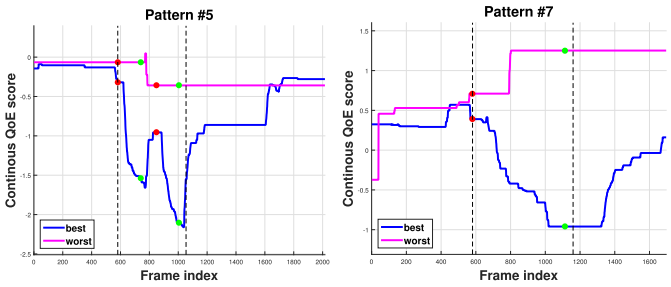


Fig. 8. Temporal ratings with the highest (blue) and the lowest (purple) degree of consistency for two playout patterns in a given video content. Left: pattern #5; Right: pattern #7. The red dots denote the start of a video impairment (rebuffer or compression) while the green dots the end of the impairment. The dashed lines mark the time interval (in frames) used in the DTW.

scale ranged from 0 (Bad) to 5 (Excellent). After each video finished, the subjects were asked to give an overall rating of their QoE using the same rating bar. Then, a screen prompt allowed the subjects to take a short break before they could initiate the playout of the next video. Examples of these steps can be seen in Fig. 7.

IV. Post Processing of Subjective Scores

A. Normalization of Subjective Scores

Following the subjective data collection, z-score normalization [3] was applied on a per session and per subject basis to account for differences in the use of the rating scale by each subject, for each of the 3 viewing sessions. Let $s_{ijk}(t)$ and f_{ijk} denote the continuous scores and the retrospective score assigned by subject i to video j during session k and let t denote the frame number. Note that the set of all j videos viewed by subject i may not have been exactly the same for another subject i'. Consider the following operations:

$$\hat{s}_{ijk}(t) = \frac{s_{ijk}(t) - \mu_{s,ik}}{\sigma_{s,ik}}, \quad \hat{f}_{ijk}(t) = \frac{f_{ijk} - \mu_{f,ik}}{\sigma_{f,ik}} \tag{1}$$

where $\mu_{s,ik}$, $\mu_{f,ik}$ are the mean continuous and retrospective scores assigned to all videos at session k of subject i and $\sigma_{s,ik}$, $\sigma_{f,ik}$ are the corresponding standard deviations. Since the generated video patterns are of different duration because of the introduction of rebuffering events, computing temporal Differential Mean Opinion Scores (DMOS) was not possible.

B. Subject Rejection Using Continuous Scores

Using the subjective data in the form of z-scores, the next step was to apply subject rejection strategies to identify potential outliers in the rating process. In video quality studies with longer videos, it is possible that subjects demonstrate less motivation and/or attention on some videos than on others. While subject rejection is not a sophisticated model of human attention, we think that it is sufficient to filter out inattentive subjective responses. In a recent work, a model of subjective consistency and bias was proposed for recovering improved subjective scores in the retrospective QoE setting [17].

We believe that subject rejection methodologies based only on retrospective scores are questionable for the following two reasons. First, if some subject is rejected based on only a single score per video but then is also discarded from all other video sequences he or she viewed (as is typically done), such a strict rejection criterion may needlessly reduce the amount of data. In our case, applying subject rejection only on the retrospective scores as suggested in [16], [3] led to 7 subjects being marked as outliers. Since we focused on the temporal effects of subjective QoE, we considered it sensible to enrich the subject rejection strategy by taking into account the temporal dimension of subjective QoE.

In our preliminary design of temporal subject rejection schemes, we experimented with simple heuristics. First, we applied the frame-to-frame equivalent of retrospective score rejection [3], [16], [18] which yielded inconsistent results. We believe this was due to the fact that introducing both dynamic bitrate changes and rebuffering events led to more complex subject reactions with different response and lag times. An alternative approach is to apply a simple thresholding method: discard subjects that are un-responsive during any rebuffering event. However, we encountered instances where subjects did not react to a rebuffering event but were very unforgiving of a second rebuffering. This observation led us to avoid using such simple *ad hoc* methods.

We instead deployed a more sophisticated dynamic time warping (DTW) [19] strategy on the subjective ratings to identify similarities in aligned *temporal* subject responses. Subjects that were completely un-responsive during a time period where most of the other subjects reacted were noted. To demonstrate the usefulness of the DTW approach to study and identify inconsistent temporal behavior among subjects, consider the examples shown in Fig. 8. Both examples depict the most and the least consistent human raters of a given video sequence, one with two rebuffers and one with a bitrate drop to 100 kbps. In the first case, it is clear that the least consistent subject did not react to any of the rebuffering events, whereas the most consistent subject had a more predictable QoE reaction. Similar behavior occurs in the second case: the subject marked with blue lowered the QoE during the bitrate drop, while the least consistent subject had a highly unreliable QoE reaction: during the bitrate drop, the recorded QoE increased.

We now define the input to the DTW. Consider subject i and the temporal rating waveform s_{ij}, where j denotes a video content using one of the 8 playout patterns. Our main focus was occurrences of rebuffering or compression events

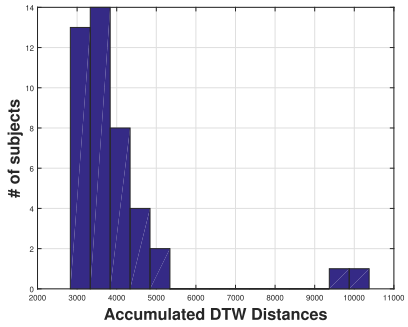


Fig. 9. Distribution of accumulated DTW distances computed on one test video. The rightmost subjects have a higher chance of being outliers.

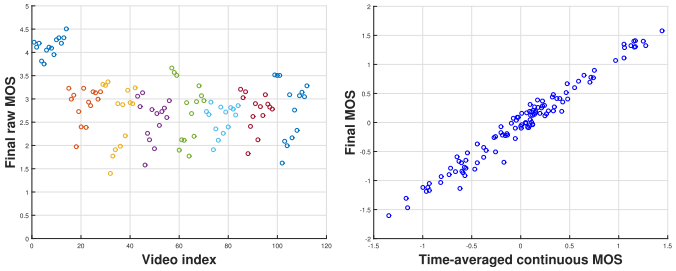


Fig. 10. a) Raw MOS for all 8 patterns. Only pattern #0 is significantly different from the other 7. b) Scatter plot of the frame-averaged continuous scores (horizontal axis) against the retrospective MOS (vertical axis) for all test videos.

since those are the key aspects that determine subjective QoE. Therefore, we trimmed the s_{ij} waveforms by selecting the time interval between the first video impairment (rebuffer or bitrate change) that took place until the last one occured. To capture the temporal behavior when normal playback (playout rate of 250 kbps) resumed, we lengthened this time interval by 4 seconds. An example of a considered time interval can be seen in Fig. 8. We set the DTW window size to be 10% of s_{ij}. Similar values of the window size ranging between 5% and 10% yielded similar results.

We collected all warped distances between subjects i and k, i.e., $d_{ik} = \text{DTW}(s_{ij}, s_{kj})$, where d_{ik} denotes the temporal misalignment between subjects i and k. This is a measure of dis-similarity: a large d_{ik} could mean that subject i reacted very rapidly to some stimuli whereas subject k reacted more slowly. Subject ratings having large distances from most of the others can be thought of as unreliable. As we have already explained, however, only per video rejection decisions were made, i.e., if subject i had unreliable ratings on some video j it did not imply rejection of all the other subject's ratings. To eliminate biases introduced by the individuality of subject scoring strategies, each subject's continuous rating waveform was linearly scaled independently to cover the range $[0,\ 1]$.

Computing the DTW warped distances, d_{ik} yielded a matrix $\mathbf{D} = [d_{ik}]$ describing the temporal misalignments between all subjects that viewed video j. Since the DTW distance is symmetric, we computed only the upper triangular part of the matrix and set the diagonal entries to 0. Then, the sum of the DTW distances across the rows (or columns) of $\mathbf{D}$ may be considered to be a measure of how unreliable a subject is: a large accumulated distance implies a subject whose responses were consistently mis-aligned with respect to other subjects when rating the same video.

In Fig. 9, the distribution of accumulated DTW distances is shown for one of the test videos. The horizontal axis corresponds to the sum of the rows in $\mathbf{D}$, while the vertical axis indicates the number of subjects having the corresponding DTW distance. The distribution of accumulated distances is skewed to the right, making outlier identification more challenging. A standard technique is to apply Tukey's boxplot [20] rule, i.e., mark all observations that are smaller than or that exceed 1.5IQR as outliers, where IQR is the interquartile range Q_3-Q_1 where Q_1 is the 25th percentile and Q_3 the 75th percentile. However, this rule assumes an underlying normal distribution. To address the skewness of the data distribution, we can either transform the data using an appropriate transformation (e.g. a Box-Cox [21] transformation) or use an adjusted boxplot technique like the one in [22]. We used the adjusted boxplot method. Then, an observation is considered to be an outlier if it lies outside the interval:

$$[Q_1 - h_l(\text{MC})\text{IQR}\ \ Q_3 - h_u(\text{MC})\text{IQR}] \tag{2}$$

where h_l and h_u are functions of the medcouple (MC), which is a skewness measure [22]. We used the exponential model proposed in [22] i.e. $h_l = 1.5\exp^{\alpha \text{MC}}$ and $h_u = 1.5\exp^{\beta \text{MC}}$, where α and β are weighting factors. We picked $\alpha = -4$ (default value) and $\beta = -1$ since the DTW distributions are right skewed, and a small value of β produced a more robust estimator. Using this skewness-driven boxplot, we identified potential outliers on each test video and removed them from the collected data.

V. Analysis of Retrospective Scores

We next discuss how we analyzed the subject scores using retrospective scores. First, we considered the overall distribution of the retrospective MOS before z-scoring. Figure 10a shows the distribution of raw retrospective MOS. It can be observed that the scores varied over the interval $[1.5,\ 4.5]$, hence the entire scale $[0,\ 5]$ was not used. However, the subjects were not prompted to use the entire scale, since this could introduce bias. Instead they were allowed to give their natural responses. Also, note that patterns #1 to #7 were given similar MOS scores, while pattern #0 was consistently rated higher by subjects (over all contents). This not surprising since this pattern assumes a rebuffering-free scenario where the encoding bitrate is a constant 500 kbps.

In typical streaming applications, subjects are exposed to long video sequences, and events that occur early on may have less effect on the overall rating given by a subject. This is known as the "recency effect" [7] where recent events more heavily influence the current perception of one's viewing experiences.

To examine these biases further, we conducted a preliminary statistical analysis to determine whether the playout patterns were actually (retrospective) scored differently by the subjects. We verified that the score distributions were not very skewed, then applied the Wilcoxon ranksum test [23] (using significance level $\alpha = 0.05$). We observed that, in many cases,

the statistical comparisons between the retrospective scores assigned to playout patterns yielded statistically insignificant differences. This could be explained by recency (latest experiences matter for retrospective evaluations) and the duration neglect effect [7]: subjects may lower their temporal scores if a long lasting video impairment occurs. However, even if they did recall the duration of an impairment, they tended to be insensitive to its duration when making retrospective QoE evaluations. Also, note that, by the time the subjects were asked to give an overall evaluation of each test video, more than 15 or 20 seconds of the 250 kbps playout had occurred. Given the tendency of subjects to evaluate videos based on more recent experiences, the test videos were possibly rated in response to the most recent video behavior.

If one is seeking a simple and direct QoE analysis, then it would seem desirable to obtain a single QoE value for each test video. Since the retrospective scores are affected by recency and duration neglect, we used simple frame averaging on the temporal scores to obtain a summary rating of each test video. Unfortunately, averaging continuous subjective scores without first applying temporal alignment does not account for the temporal QoE behavior of each subject (such as subject response delays). However, the DTW is appropriate only for pairwise time-series alignment, and may not produce an output having the same duration as the original waveforms. In our search for a recency-insensitive summary rating, we found that simple averaging correlated well with the retrospective scores, as seen in Fig. 10b. This observation aligns with two previous subjective studies: one where the test videos lasted only 10 seconds [3] and one with longer videos [10]. Apart from frame averaging, we were also interested in explicitly capturing the subjective responses due to the impairments caused by the available bandwidth drop. In order to study those time intervals where the only visual impairments were due to the available bandwidth drop, we applied the following protocols: on patterns #1, #4, #5, #6 and #7, we applied averaging on all frames after the available bandwidth drop occurs. By contrast, on patterns #2 and #3, where there was heavier compression even prior to the bandwidth drop, all the frames were used. Since pattern #0 was impairment-free, we did not include it in the comparisons.

Using the averaged scores as the summary ratings, we compared the playout patterns of each content as shown in Fig. 11. Clearly, the ratings given to patterns #5 and #6, which belong to the second category, where no buffer was utilized, were statistically inferior to those of the patterns from the first category (#1 to #4), since the available buffer was zero and fewer bits were spent; hence there were more frozen frames due to rebuffering events and/or lower bitrate values. By comparing pattern #4 with #2 we observed that a consistently low bitrate value (to avoid rebuffering), as in the "conservative" client strategy #2, was not tolerated by subjects. Further, subjects preferred a long rebuffering (#1) if it meant better quality elsewhere rather than the combination of a short rebuffering event combined with an intermediate recovery bitrate (#3).

An important aspect of the interactions between rebuffering and compression is whether there exists a "compression threshold", i.e., a bitrate level below which rebuffering will

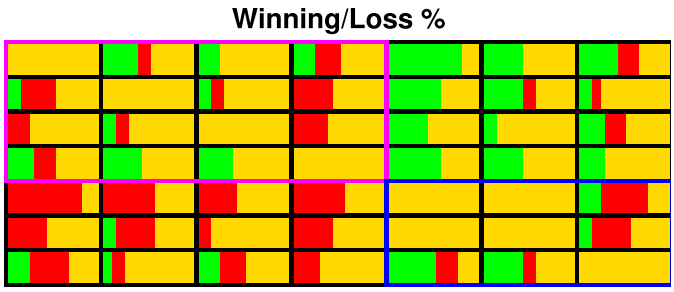


Fig. 11. Wilcoxon ranksum test using $\alpha = 0.05$ on the averaged temporal scores for all patterns, represented as a 7×7 matrix. Each entry shows the winning percentage of the row compared to the column for all 14 video contents. Green shows the number of contents that the pattern in the row is QoE superior to the column, red shows the contents where the row is inferior to the column and orange shows that the row and column are indistinguishable. The purple box shows the comparisons only between patterns #1 to #4 ($B_0 = 1333$ kbits) and the blue box shows the comparisons only between patterns #5 to #7 ($B_0 = 0$ kbits).

be preferred over a highly compressed stream. Clearly, this "threshold" may be different across contents depending on the content's spatio-temporal complexity. Here, we can perform such a comparison directly, since both playback states (normal playback at a much lower bitrate as in #4 and playback interruption as in #1) are equalized in terms of bandwidth usage.

By comparing rebuffering with transient bitrate drops (see first row and fourth column of Fig. 11) we found that the outcome of the statistical comparison depended on the level of content complexity. Out of the 14 contents, subjects preferred a very low bitrate in 4 of them, rebuffering in 3 and for the remaining 7, the statistical test yielded a statistical equivalence betwen #1 and #4. Notably, all 4 contents where subjects preferred #4 were slow motion scenes (e.g. a dialogue between actors) and/or low spatial complexity scenes, while the 3 contents where rebuffering was preferred were contents of either high spatial complexity (as in the ElFuente fountain scene) or high temporal complexity (e.g. a fight scene rich in motion); hence they required more bits to be encoded. This observation strongly highlights the trade-off between rebuffering and compression artifacts in perceived QoE.

Notably, pattern #7 had the best performance among patterns in the second category ($B_0 = 0$) and was comparable to #2 and #3. Again, this shows that subjects preferred transient bitrate drops. Surprisingly, #7 used fewer bits than #2 and #3 but yielded similar QoE. While #7 assumed an ideal client that could immediately adapt to the network conditions, this comparison demonstrates the merits of QoE-aware network policies: using fewer bits does not always mean that perceived quality is lower. However, we also observed that patterns #5 and #6 were statistically indistinguishable over all contents. This brings up another aspect of the subjective test's design: apart from recency, allocating the same number of bits under these circumstances could signify a similar retrospective QoE or summary rating. This underlines the need to exploit the temporal aspects of QoE, since retrospective ratings reveal only some aspects of subject QoE.

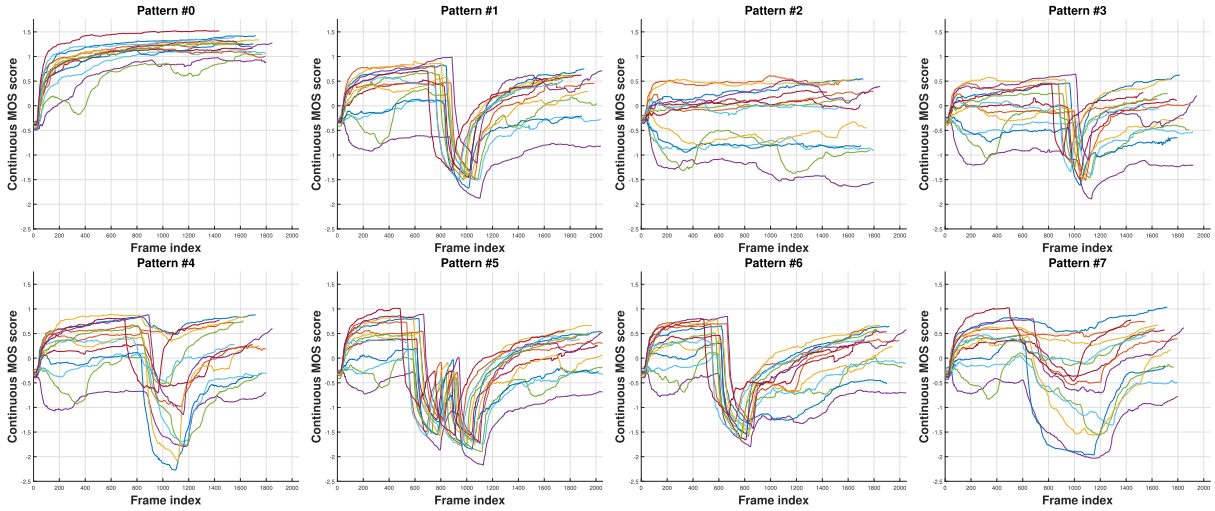


Fig. 12. Temporal ratings across all contents for all playout patterns after subject rejection. First row: patterns 0 to 3; second row: patterns 4 to 7.

VI. Analysis of Temporal Scores

Temporal scores are a rich source of subjective QoE. Similar to the frame averaging used earlier, we performed frame averaging on the continuous subjective scores and show the result for several patterns in Fig. 12. We now focus on a comparison between patterns #1 and #7. Clearly, rebuffering (#1) severely and sharply damages subjective QoE for all contents. Further, the QoE recovers at a slower pace than it originally dropped, suggestive of the hysteresis phenomenon: there is a lag between subjective QoE scores and current video quality or playback status. We earlier observed that subjects were not forgiving of rebuffering events. By contrast, when the bitrate dropped from 250 to 100 kbps, the subjective QoE reactions varied depending on each content. On scenes having higher spatiotemporal complexities, compression artifacts may be more visible and affect the QoE heavily and sharply, while others may not be affected to the same extent. Similar observations may be made for all patterns that contain at least one rebuffering event (where the video freezes and the rebuffering icon appears), which are obvious and unpleasant to viewers, whereas bitrate drops have a different impact on subjective QoE depending on each content's complexity.

Notably, the constant encoding bitrate employed in #2 had a temporally varying effect on the perceived QoE. Given the long duration of the video contents and the different video characteristics present in each content (such as scene changes), it is clear that the subjects' QoE also changed over time even when the encoding scheme was static. This observation strongly supports a "per chunk" encoding strategy [24], where each video content is first split into short video chunks and then, based on the video complexity during this chunk, an appropriate encoding scheme can be chosen.

To investigate the interplay between rebuffering and compression artifacts under a different light, we split the test contents into two sets based on their complexity: Set 1 includes source contents of low complexity and Set 2 those of higher complexity. To determine the two sets we considered the following: contents with high motion and/or spatial complexity require more encoding bits, hence subjective scores would likely be lower on such sequences. To determine content complexity, the authors of [25] defined a criticality measure as the logarithm of the sum of the SI and TI indices.

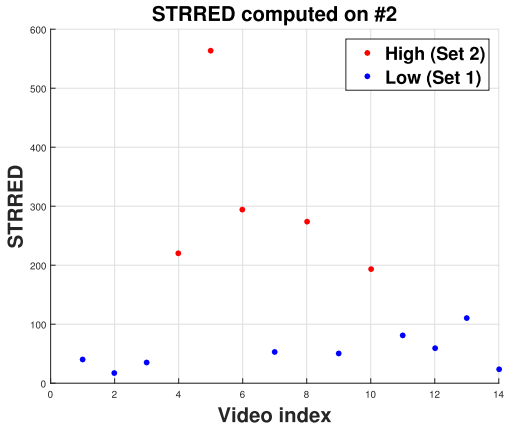


Fig. 13. ST-RRED values between pattern #2 and the original source video for all 14 contents. Blue points correspond to low complexity contents while red points correspond to high complexity contents.

Given that the quality impairments of the otherwise very high quality videos being viewed are dominated by H.264 compression, an excellent measure of the content complexity to a fixed bitrate are the scores of a high performance objective quality engine such as ST-RRED [26]. ST-RRED is an information-theoretic approach to VQA that builds on the innovations in [27] and [28]. It achieves quality prediction efficiency without the need to compute motion vectors unlike [29] and [30].

To avoid any subjective biases due to content, we computed ST-RRED [26] between the original pristine video and #2 - constant encoding bitrate under the same total bit budget constraint. The computed ST-RRED value (on the constant bitrate encodes) was a way of describing the content complexity: the higher the ST-RRED value, the less "complex" the content was assumed to be. As we will show later, ST-RRED

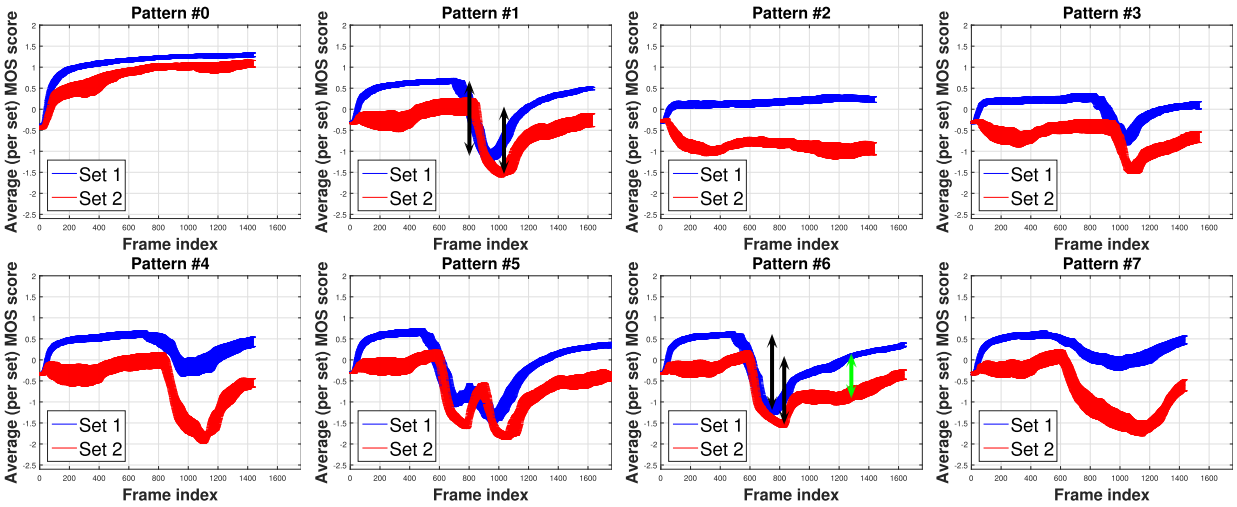


Fig. 14. Averaged temporal ratings and standard errors for content Sets 1 and 2 for all playout patterns after subject rejection. First row: patterns 0 to 3; second row: patterns 4 to 7. Due to the different video lengths, we trimmed the axis of the plot to the duration of the shortest video sequence. The black arrows show the effect of rebuffering for the high vs. low complexity sets. The green arrow shows the different rates of QoE recovery for these sets.

performed the best among the VQA models studied across the subset of video sequences without any rebuffering, hence it was deemed suitable for this purpose. Finally, as shown in Fig. 13, there are 5 contents (shown in red color) that have a relatively higher encoding complexity than the rest. Therefore, we considered those 5 as Set 2 while the rest were assigned to Set 1.

Next, we found the average (per frame) MOS score over all contents for each of the 8 different patterns, as shown in Fig. 14. The effects of content complexity were evident: after a rebuffering event occurred, the QoE recovered more slowly for contents in Set 2 (high complexity) as shown by the green arrow in #6. Meanwhile, the videos in Set 2 tended to have larger standard errors against the videos in Set 1, since the increased encoding complexity may have led to a larger variance in the subjective QoE reactions. Overall, during normal playback, the contents in Set 2 have a lower QoE than the contents in Set 1.

We also observed the following interaction: a relatively long rebuffer event (as in playout patterns #1 and #6) led to larger drops in the reported subjective QoE on Set 1, as compared to Set 2 (see the black arrows in the plots for playout patterns #1 and #6). It is likely that the subjects were more annoyed by rebuffering events when they occurred during the playback of higher quality video content. A similar observation was also made in [6] using retrospective QoE ratings on short video sequences. However, for shorter rebuffering events (playout patterns #3 and #5) quality drops due to rebuffering between the two sets was similar. Notably, the second rebuffering in pattern #5 led to the opposite effect: given that one rebuffering event had already occurred, the quality drop on Set 2 was larger than the one for Set 1. This may be attributed to the effects of memory of a recent rebuffering event on currently perceived QoE.

By comparing patterns #1, #3 and #5, it is also evident that when the number or the durations of the rebuffering events increases, there is a larger drop in the temporal QoE scores. Again, these effects of rebuffering on the subjective QoE were harder to capture when we used the retrospective QoE ratings.

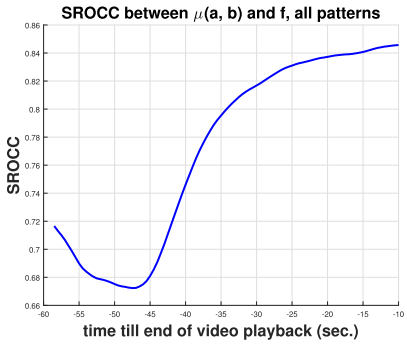


Fig. 15. SROCC between the averaged temporal scores (over a 10 sec. window) and the retrospective MOS.

VII. Cognitive Aspects in Subjective QoE

A. Recency Effects

As already discussed, subject QoE might depend heavily on more recent experiences. To further investigate this claim, we performed local averaging on the temporal scores using a sliding window, then measured the correlations of those averages against the retrospective scores. Let κ denote the size of the sliding window in seconds, τ be the total duration of a video and $\mu\ (a,\ b)$ be the average of the temporal scores from frame a to frame b and f be the retrospective score assigned to that video. Figure 15 shows the SROCC between $\mu\ (a,\ b)$ and f using $\kappa = 10$ seconds. It is clear that local temporal averaging produced stronger correlations over the more recent time intervals. This agrees strongly with the recency effect observed on the subjects' QoE.

B. Non-Linearities in Subjective QoE

Non-linearities in human responses to video quality are usually not considered in depth. Here, we are able to examine

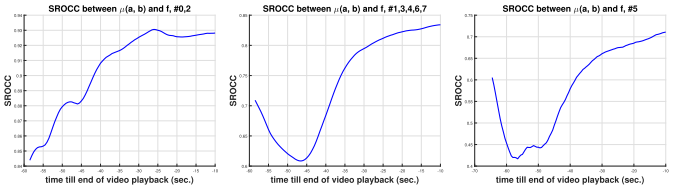


Fig. 16. Spearman's rank correlation coefficient for different pattern sets. From left to right: patterns 0 and 2, patterns 1, 3, 4, 6, 7 and pattern 5.

these effects given the richness of the collected temporal data. Fig. 15 shows that, as the observation window is increased further into the past, the rank correlation decreases until approximately 45 seconds, at which point it increases. This could be due to the fact that after the first 15 seconds most of the video impairments begin to occur, hence a local temporal window of "high disagreement" between subjects occurs as the impairments take place. By high disagreement, we refer to different response times between subjects, different recovery times and different use of the rating scale. Note that even after z-scoring normalization, the subject ratings are still dependent on the rating behavior over time. We refer to both bitrate changes and rebuffering events during those time intervals as "events" where non-linearities in the human responses are activated and intensified. As a result, linearly combining the scores still produces non-linear measurements that do not correlate as well as when such events are not taking place.

To examine our hypothesis, we considered three different cases in Fig. 16: when the encoding bitrate is constant (patterns #0 and #2), when there is a single event (or two consecutive events) such as a lone rebuffering event or one followed by a bitrate drop (patterns #1, #3, #4, #6 and #7) and when there are two distinct events (pattern #5). The first case demonstrates the recency effect: more recent scores correlate more highly with the retrospective score. In the second case, a combination of recency and human non-linearities is demonstrated: past experiences correlate less with the retrospective score, especially when there is a bitrate drop or a rebuffering event. However, recency by itself is not enough to explain subject QoE when a very negative QoE experience has occurred in the past. As shown in the third case, the correlation is much lower even over very recent time intervals due to the two rebuffering events that have happened earlier.

C. Recency Versus Primacy

The previous analysis gives rise to the following contradiction: if subjects tend to bias their ratings based on the recency effect, why would a rebuffering event (or a bitrate drop) that happened much earlier matter? In the cognitive science literature, the *primacy* effect refers to the human tendency to recall events that occurred at the beginning of a series of events [31]. We can apply this concept to the various events to which subjects are exposed when viewing streaming videos, such as bitrate changes. It is likely that, when giving a retrospective evaluation, both primacy and recency effects affect the subjects' responses. If the perceived video quality is relatively stable, subjects tend to internally rely more heavily on their latest experiences to make a retrospective decision, yet negative QoE events that occur early on can also activate longer term reactions.

TABLE I

SPEARMAN'S RANK CORRELATION COEFFICIENT (SROCC) FOR VARIOUS IMAGE/VIDEO QUALITY ASSESSMENT ALGORITHMS (IQA/VQA) AGAINST THE RETROSPECTIVE SCORES AFTER PERFORMING MEAN POOLING ON THE NO-REBUFFERING SUBSET (S_q) AND ON THE WHOLE DATASET (S_{all}). THE BEST RESULT PER SUBSET IS IN BOLDFACE

IQA/VQA metric	S_q	S_{all}
PSNR (IQA, FR)	0.5535	0.5257
PSNRhvs [34] (IQA, FR)	0.5884	0.5465
SSIM [35] (IQA, FR)	0.7862	0.7230
MS-SSIM [36] (IQA, FR)	0.7647	0.6979
NIQE [37] (IQA, NR)	0.3811	0.1300
VMAF [38] (VQA, FR)	0.7607	0.6079
ST-RRED [26] (VQA, RR)	**0.8216**	**0.7257**
GMSD [39] (IQA, FR)	0.6665	0.5937

Given our previous analysis of both retrospective and temporal scores, it is important to summarize the different aspects of each. For long video sequences in streaming applications, retrospective scores are simple and efficient QoE descriptors but do not capture all aspects of QoE. When integrating their temporal experiences into a single QoE score, subjects may be biased towards recent experiences (recency) or much earlier but memorable - typically unpleasant - ones (primacy), but they may also be insensitive to how long these unpleasant viewing experiences were (duration neglect). By contrast, temporal scores are rich and descriptive QoE indicators. However, the different response times between subjects and other temporally varying QoE aspects make temporal scores harder to analyze.

VIII. OBJECTIVE VIDEO QUALITY ASSESSMENT

A. Is Objective VQA Enough?

Most VQA algorithms are not applicable to frame freezes; hence video sequences with playback interruptions are usually not considered in objective quality analysis studies [4]. As a way of understanding how well these "standard" VQA models predict subjective QoE, we ask the question: "How well do VQA algorithms perform on video sequences when excluding frame freezes?" To answer this question, we considered the set S_q of videos without any rebuffering, the set S_r of videos having at least one rebuffering event and the whole dataset ($S_{all} = S_q \bigcup S_r$). Clearly, S_r and S_q are disjoint. Then, we applied various quality metrics on S_q and S_{all}. We compared several leading full reference (FR), reduced reference (RR) and no reference (NR) image (IQA) or video (VQA) quality assessment algorithms [32], [33]: PSNR, PSNRhvs [34], SSIM [35], MS-SSIM [36], NIQE [37], VMAF [38], ST-RRED [26] and GMSD [39]. We refer the interested reader to [30], [40], and [41] for other perceptual VQA models that have been developed. When applying them on the videos in S_q, we calculated the quality scores only on normal playback frames and measured the correlation with the retrospective

TABLE II

a) SROCC AGAINST THE RETROSPECTIVE SCORES ACHIEVED WHEN USING TEMPORAL POOLING STRATEGIES ON THE LIVE-NETFLIX DATASET, SETS S_q AND S_{all}. FOR EACH QUALITY METRIC AND SUBSET (S_q/S_{all}), THE BEST POOLING METHOD IS IN BOLDFACE. THE BEST COMBINATION (QUALITY MODEL AND POOLING) PER SUBSET IS IN BOLDFACE AND ITALIC FONT. b) MEDIAN SROCC AND LCC AGAINST THE RETROSPECTIVE SCORES ON S_{all} FOR DIFFERENT QoE MODELS AFTER PERFORMING 1000 TRAIN AND TEST TRIALS. THE BEST RESULT IS DENOTED BY BOLDFACE

(a)

Set	S_q				S_{all}			
Model/Pooling	mean	hysteresis	VQ	percentile	mean	hysteresis	VQ	percentile
PSNR	0.5535	0.5518	0.5621	**0.5869**	0.5257	0.5360	0.5398	**0.5581**
PSNRhvs [34]	0.5884	0.5960	0.6134	**0.6379**	0.5465	0.5601	0.5668	**0.5781**
SSIM [35]	0.7862	0.7971	0.7899	**0.8049**	0.7230	**0.7298**	0.7028	0.7051
MS-SSIM [36]	0.7647	0.7686	0.7593	**0.7800**	0.6979	**0.7037**	0.6680	0.6772
NIQE [37]	0.3811	0.4094	**0.4185**	0.2720	0.1300	0.1412	**0.1590**	0.0110
VMAF [38]	0.7607	**0.7760**	0.7679	0.7663	0.6079	**0.6347**	0.6116	0.5006
ST-RRED [26]	0.8216	0.8154	0.8032	**0.8232**	**0.7257**	0.7235	0.7139	0.7174
GMSD [39]	0.6665	0.6465	0.6416	**0.7502**	0.5937	0.5634	0.5684	**0.6369**

(b)

Type	Method	SROCC	LCC
rebuffering	FTW	0.3403	0.2956
	VsQM	0.3120	0.2421
IQA/VQA	PSNR	0.5983	0.5693
	SSIM	0.6843	0.7487
	MS-SSIM	0.6791	0.7286
	NIQE	0.1852	0.4051
	VMAF	0.5983	0.7497
	ST-RRED	0.6791	0.7543
	GMSD	0.6470	0.6957
hybrid	PSNR+SQI	0.5574	0.5937
	SSIM+SQI	0.7583	0.8095
	MS-SSIM+SQI	0.7470	0.7851
	PSNR+ATLAS	0.6743	0.8137
	SSIM+ATLAS	0.8714	0.9272
	MS-SSIM+ATLAS	0.8522	0.9088
	ST-RRED+ATLAS	**0.8848**	**0.9364**

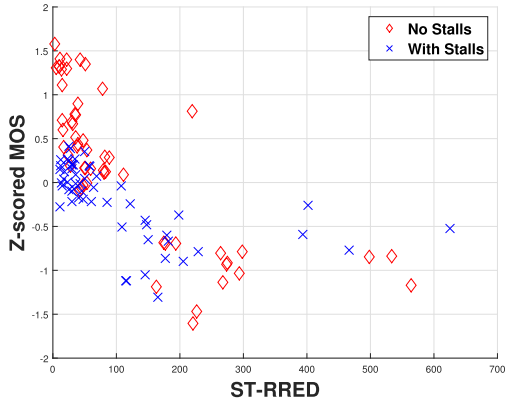


Fig. 17. Performance of ST-RRED on videos with and without rebuffering.

scores after subject rejection. For PSNRhvs we used the publicly available Daala [42] implementation and for the other methods we used the available implementations. All models were applied on the luminance channel of the test videos and the black borders around the videos were removed. The results are presented in Table I.

As shown in the first column, NIQE unsurprisingly performed the worst since it is a frame-based NR model, while PSNR and PSNRhvs performed the worst across all FR algorithms, followed by GMSD. The results on S_{all} were much lower than on S_q; indicating that the tested IQA/VQA systems were unable to predict QoE as well when rebuffering events were present. Note that SSIM performed better than MS-SSIM and close to the best predictor (ST-RRED) on S_{all}. This suggests that the subjects were internally responding strongly to rebuffering events rather than evaluating quality only. To investigate the performance of VQA models on videos afflicted by rebuffering, Fig. 17 shows the performance of ST-RRED on videos with rebuffering and on videos without rebuffering coded by color and symbol shape. It is important to observe that the predictive power of ST-RRED decreases when rebuffering events are introduced, which is not surprising: almost all perceptual IQA/VQA models only consider the effects of visual quality on the perceived QoE. Therefore, in the presence of rebuffering, objective video quality models become less reliable predictors of subjective QoE. This implies the need to develop more general QoE-aware methods.

B. Temporal Pooling Strategies for Objective VQA

Simple averaging of frame quality scores is broadly used to pool quality scores computed on short videos, but more sophisticated perception-driven temporal pooling strategies have been proposed, including hysteresis [43], VQ pooling [44] and temporal percentile pooling [45]. For percentile pooling, we sorted the frame-based values, then averaged the 5% of them which corresponded to lowest quality. We next investigated whether adopting these approaches could produce better correlations against human subjective scores on S_q and S_{all}. The results of this experiment are presented in Table IIa.

On S_q, most methods benefited from a temporal pooling strategy, except ST-RRED (where performance was improved only slightly by percentile pooling). Otherwise, these improvements were not significant for S_{all}. This suggests that deploying temporal strategies designed for short sequences may not significantly improve QoE prediction on long sequences suffering from both rebuffering and bitrate changes: temporal pooling strategies operate on the numerical scores produced by objective video quality models. Again, ST-RRED performed the best on S_q in terms of SROCC. On S_{all}, SSIM (with hysteresis pooling) was able to reach the maximum predictive performance of ST-RRED. Notably, percentile pooling was beneficial to FR methods such as SSIM and MS-SSIM on S_q, but in the case of NIQE, the prediction performance dropped considerably. This is likely because NIQE is frame-based, does not capture temporal information critical to QoE prediction, does not capture artifact fluctuations nor does it benefit from reference information. Therefore, NIQE scores may unreliably reach extreme values.

C. VQA, Rebuffering-Aware and General QoE-Aware Models

Predicting subject QoE in the presence of both rebuffering and bitrate excursions is a hard problem. One limiting

factor is that most databases do not contain both compression and rebuffering events, hence are not adequate for modeling streaming applications where both frequently occur together. The LIVE-Netflix dataset allowed us to consider these two effects together and to compare different QoE prediction models: VQA models, rebuffering-aware models such as the VsQM model proposed in [46], FTW [47], SQI [6] and Video ATLAS [48]. Thus unlike [49] and [50], we compared methods that combine perceptual VQA models with rebuffering-aware information. In the objective evaluations, we only considered the retrospective scores.

Since Video ATLAS is a learning-based approach, we performed 1000 trials using 80% train and 20% test content splits to avoid content biases. We evaluated Video ATLAS using different regression models and reported the best performing regressor. For SQI, we determined the best parameters using the training contents for each trial. We report the median SROCC and LCC values over all 1000 trials in Table IIb. To compute LCC, we first applied a non-linear (logistic) regression on the output QoE scores as suggested in [16]. It may be observed that embedding rebuffering-aware information into the IQA/VQA models produced significantly improved performance.

IX. Discussion and Conclusion

We described a subjective study that focused on the temporal aspects of subjective video QoE under various network, buffer and low bitrate constraints. The study gathered both continuous time and retrospective data that we processed to extract useful information regarding those factors that affect QoE, such as the network condition, the encoding bitrate and the spatio-temporal complexities of the video contents being viewed. Overall, we hope that QoE researchers find the new database to be a useful tool for studying the temporal aspects of subjective quality of experience. This remains a relatively unexplored area of research that poses many challenges.

We plan to continue studying the various aspects of human responses when viewing videos streamed under realistic network conditions since better models of these responses could greatly benefit future efforts to improve network streaming and encoding strategies adopted by content providers. Objective prediction models that incorporate spatio-temporal aspects of videos and that predict human reactions to both bitrate dynamics and rebuffering events could ultimately help streaming video companies address resource allocation problems more efficiently and in a user-adaptive way. Recent efforts on the overall (endpoint) QoE prediction problem [6], [48], as well as on continuous-time QoE prediction [18], [51]–[53] are important early steps towards this research goal. In the future, we plan to extend our work by focusing on continuous time QoE monitoring.

Appendix

A. Explaining the Playout Pattern Parameters

We provide an example of how some of the playout pattern parameters were determined. We fixed the rebuffer duration for pattern #1 (see Fig. 3) to 8 sec. and the average bitrate for the client in pattern #2 to be $R_2 = 160$ kbps. Since there is no rebuffering event in pattern #2 but the available bandwidth is 100 kbps for d seconds, the client in #2 expends all of the available buffer B_0 in d seconds hence $(R_2 - 100)d = B_0$ yielding $B_0 = 1333$ kbits. Let t_b be the time interval after the available bandwidth drops until a rebuffering event occurs in #1. Clearly, $t_b(250 - 100) = B_0$ since the client depletes all of the buffer before the playback interruption. During the rebuffering event, the buffer fills to $B_1 = 800$ kbits in 8 seconds, given the available bandwidth of 100 kbps. The client chooses to start the playback t_a seconds before the available bandwidth recovers hence $t_a(250 - 100) = B_1$, since we assume that all playout patterns eventually deplete the entire buffer. Therefore, $t_a = 5.3333$ sec. and $d = t_e + 8 + t_a \approx 22.2167$ seconds.

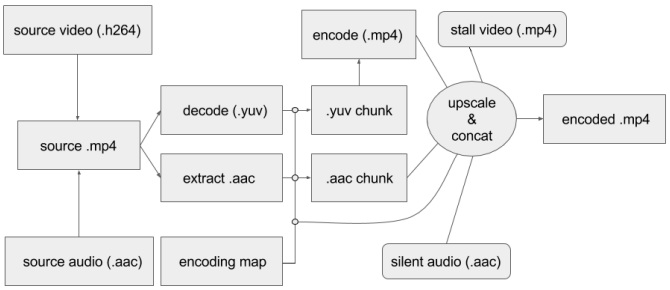


Fig. 18. Encoding pipeline used to create the playout patterns.

B. Implementation Details of the Encoding Pipeline

Each high quality video source sequence is first encoded into H.264 format, combined with a corresponding, synchronized audio stream and placed in an mp4 container without further re-encoding. Then, following the application of a specific network-simulated pattern, the .mp4 file is divided into a number of different chunks, each at a different encoding bitrate. For example, pattern #6, which contains both bitrate changes and a rebuffering event would have three chunks: one for the rebuffering event and two corresponding to the encoded video before and after the rebuffering event.

The encoding pipeline then assembles the segments of the final video, by concatenating them using an encoding profile demarking the interval of time spent at each quality level. The location and duration of each rebuffering event is specified as: enc $<$ start $>$ $<$ stop $>$ $<$ bitrate $>$ stall $<$ start $>$ $<$ duration $>$, with time measured in seconds and bitrate in kbps. The encoding resolution was based on the used bitrate and the encoding profile was set to high.

Using this encoding profile, the encoding process was carried out as follows (see Fig. 18). First, the source video and audio streams were transferred from Google Drive and stored locally for further encoding. Next, the source video stream (in H.264 format) was decoded, yielding an uncompressed raw .yuv file. The encoding map was then used to split the .yuv file in a frame-accurate manner, yielding .yuv chunks, e.g. three chunks for pattern #6. A two pass encoding step using FFMPEG was then applied to encode the .yuv files into .mp4 format. For pattern #6, this corresponds to two chunks

encoded at 250 kbps, and one encoded at 160 kbps. The final frame of every video chunk that occurs immediately before a rebuffering event was used to generate a "rebuffering video chunk". A familiar "loading icon", (a spinning wheel) was overlaid on that frame during the rebuffering event and animated to simulate the desired video rebuffering effect. After encoding each of the yuv chunks into .mp4 format, all of the .mp4 segments were upscaled to the device resolution (1080p), then concatenated into a single .mp4 file. For playback purposes, each concatenated .mp4 file was lightly compressed using CRF 10, since raw playback on mobile devices is not supported.

ACKNOWLEDGEMENT

The authors would like to acknowledge Te-Yuan Huang and Maria Kazandjieva of Netflix for their help in designing the playout patterns.

REFERENCES

[1] Cisco Corp. (2011). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2010–2015*. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 187–198, 2015.

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

[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] J. Y. Lin, R. Song, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *J. Vis. Commun. Image Represent.*, vol. 30, pp. 1–9, Jul. 2015.

[6] 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.

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

[8] N. Staelens *et al.*, "Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 707–714, Dec. 2014.

[9] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant, "Study of the effects of stalling events on the quality of experience of mobile streaming videos," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 989–993.

[10] 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. 5th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 52–57.

[11] (2017). *Consumer Digital Video Library*. [Online]. Available: http://www.cdvl.org

[12] (2017). *FFMPEG*. [Online]. Available: https://www.ffmpeg.org/

[13] I. E. Richardson, *The H.264 Advanced Video Compression Standard*. New York, NY, USA: Wiley, 2010.

[14] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[15] *ITU-T Recommendation: Subjective Video Quality Assessment Methods for Multimedia Applications*, document P.910: 2008.

[16] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document BT-500-13, Int. Telecommun. Union, Geneva, Switzerland, 2012.

[17] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Proc. Data Compress. Conf.*, Apr. 2017, pp. 52–61.

[18] 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.

[19] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. KDD Workshop*, 1994, vol. 10. no. 16, pp. 359–370.

[20] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, Jan. 1977.

[21] R. M. Sakia, "The Box-Cox transformation technique: A review," *J. Roy. Statist. Soc. D, Statist.*, vol. 41, no. 2, pp. 169–178, 1992.

[22] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Comput. Statist. Data Anal.*, vol. 52, no. 12, pp. 5186–5201, 2008.

[23] S. Siegel, *Nonparametric Statistics for the Behavioral Sciences*. New York, NY, USA: McGraw-Hill, 1956.

[24] J. De Cock, Z. Li, M. Manohara, and A. Aaron, "Complexity-based consistent-quality encoding in the cloud," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1484–1488.

[25] C. Fenimore, J. Libert, and S. Wolf, "Perceptual effects of noise in digital video compression," *SMPTE J.*, vol. 109, no. 3, pp. 178–187, Mar. 2000.

[26] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2012.

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

[28] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop Video Process. Quality Metrics Consum. Electron.*, 2005, pp. 23–25.

[29] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Apr. 2007, pp. I-869–I-872.

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

[31] A. J. Greene, C. Prepscius, and W. B. Levy, "Primacy versus recency in a quantitative model: Activity is the critical distinction," *Learn. Memory*, vol. 7, no. 1, pp. 48–57, 2000.

[32] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[33] A. K. Moorthy and A. C. Bovik, "Visual quality assessment algorithms: What does the future hold?" *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 675–696, 2011.

[34] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," presented at the 3rd Int. Workshop Video Process. Quality Metrics, vol. 4. San Francisco, CA, USA, 2007.

[35] 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.

[36] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Conf. Rec. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

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

[38] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. (Jul. 24, 2017). *Toward a Practical Perceptual Video Quality Metric*. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[39] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[40] M. H. Pinson, L. K. Choi, and A. C. Bovik, "Temporal video quality model accounting for variable frame delay distortions," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 637–649, Dec. 2014.

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

[42] (2017). *Daala Codec*. [Online]. Available: https://git.xiph.org/daala.git/

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

[44] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, Feb. 2013.

[45] 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.
[46] D. Z. Rodríguez, J. Abrahão, D. C. Begazo, R. L. Rosa, and G. Bressan, "Quality metric to assess video streaming service over TCP considering temporal location of pauses," *IEEE Trans. Consum. Electron.*, vol. 58, no. 3, pp. 985–992, Aug. 2012.
[47] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Proc. Int. Symp. Multimedia*, Dec. 2011, pp. 494–499.
[48] C. G. Bampis and A. C. Bovik, "Learning to predict streaming video QoE: Distortions, rebuffering and memory," *Trans. Image Process.*, submitted for review.
[49] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC," in *Proc. IEEE Consum. Commun. Netw. Conf.*, Jan. 2012, pp. 127–131.
[50] J. Xue, D.-Q. Zhang, H. Yu, and C. W. Chen, "Assessing quality of experience for adaptive HTTP video streaming," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2014, pp. 1–6.
[51] 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.
[52] C. G. Bampis, Z. Li, I. Katsavounidis, and A. C. Bovik, "Recurrent and dynamic networks that predict streaming video quality of experience," *Trans. Image Process.*, submitted for review.
[53] C. G. Bampis and A. C. Bovik. (Jul. 2017). "An augmented autoregressive approach to HTTP video stream quality prediction." [Online]. Available: https://arxiv.org/abs/1707.02709

Christos George Bampis received the M.Eng.Diploma degree in EE from the National Technical University of Athens in 2014. He is currently pursuing the Ph.D. degree with the Laboratory for Image and Video Engineering, The University of Texas at Austin. His research interests include image and video quality assessment, and quality of experience in adaptive video streaming.

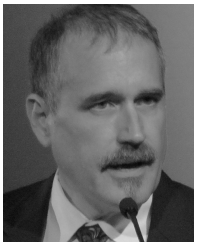
Zhi Li received the B.Eng. and M.Eng. degrees in electrical engineering from the National University of Singapore, Singapore, in 2005 and 2007, respectively, and the Ph.D. degree from Stanford University, Stanford, CA, USA, in 2012. His doctoral work focused on source and channel coding methods in multimedia networking problems. He is currently with the Video Algorithms Group, Netflix. His current interests focus on improving video streaming experience for consumers through understanding how human perceives video quality and applying that knowledge to encoding/streaming system design and optimization. He has broad interests in applying mathematics in solving real-world engineering problems. He was a recipient of the Best Student Paper Award from IEEE ICME 2007 for his work on cryptographic watermarking, and a co-recipient of Multimedia Communications Best Paper Award from the IEEE Communications Society in 2008 for a work on multimedia authentication.

Anush Krishna Moorthy received the B.E. degree in electronics and telecommunication from the University of Pune, Pune, India, in 2007, and the M.S. degree in electrical engineering and the Ph.D. degree from The University of Texas at Austin in 2009 and 2012, respectively. He joined the Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin, in 2008. He was the Assistant Director of LIVE from 2008 to 2012. He was an Advanced Imagining Engineer with Texas Instruments, USA, from 2012 to 2013, and a Senior Video Systems Engineer with Qualcomm, Inc., USA, from 2013 to 2016. He is currently a Senior Video Software Engineer with Netflix Inc., USA. His research interests include image and video quality assessment, image and video compression, and computational vision.

Ioannis Katsavounidis received the Diploma (B.S./M.S.) degree from the Aristotle University of Thessaloniki, Greece, in 1991, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1992 and 1998, respectively, all in electrical engineering. From 1996 to 2000, he was an Engineer with the High-Energy Physics Department, California Institute of Technology, Italy. From 2000 to 2007, he was with InterVideo, Inc., Fremont, CA, USA, as the Director of Software for Advanced Technologies, in charge of MPEG2, MPEG4, and H.264 video codec development. From 2007 to 2008, he served as the CTO of Cidana, Shanghai, China, a mobile multimedia software company, covering all aspects of DTV standards and codecs. From 2008 to 2015, he was an Associate Professor with the Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece, where he was involved in teaching undergraduate and graduate courses in signals, controls, image processing, video compression, and information theory. He is currently a Senior Research Scientist with Netflix, where he was involved in video quality and video codec optimization problems. His research interests include image and video quality, compression and processing, information theory, and software–hardware optimization of multimedia applications.

Anne Aaron is currently the Director of the Video Algorithms with Netflix, and leads the team responsible for video analysis, processing, and encoding in the Netflix cloud-based media pipeline. The team is tasked with generating the best quality video streams for more than 100 million Netflix members worldwide. It is also actively involved in defining next generation video through academic research collaboration and standardization work. Prior to Netflix, he had technical lead roles at Cisco, where he was involved in the software deployed with millions of Flip Video cameras, Dyyno, an early stage startup, which developed a real-time peer-to-peer video distribution system, and Modulus Video, a broadcast video encoder company.

She was born in Manila, Philippines. She received the B.S. degrees in physics and computer engineering from Ateneo de Manila University and the M.S. and Ph.D. degrees in electrical engineering from Stanford University. During her Ph.D. studies at Stanford University, she was a member of the Image, Video and Multimedia Systems Laboratory, led by Prof. B. Girod. Her research was one of the pioneering works in the sub-field of distributed video coding.



Alan Conrad Bovik (F'96) is currently a Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. His books include the *Handbook of Image and Video Processing*, *Modern Image Quality Assessment*, and *The Essential Guides to Image and Video Processing*. He received many major international awards, including the 2017 Edwin H. Land Medal from the Optical Society of America, the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, and the 2013 Society Award from the IEEE Signal Processing Society. He is a fellow of the Optical Society of America and SPIE. He co-founded and was the longest serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING and created the IEEE International Conference on Image Processing in Austin, Texas, in 1994.