Comparative Evaluation of User Perceived Quality Assessment of Design Strategies for HTTP-Based Adaptive Streaming

Ayush Bhargava, James Martin, and Sabarish V. Babu, Clemson University

HTTP-based Adaptive Streaming (HAS) is the dominant Internet video streaming application. One specific HAS approach, Dynamic Adaptive Streaming over HTTP (DASH), is of particular interest as it is a widely deployed, standardized implementation. Prior academic research has focused on networking and protocol issues, and has contributed an accepted understanding of the performance and possible performance issues in large deployment scenarios. Our work extends the current understanding of HAS by focusing directly on the impacts of choice of the video quality adaptation algorithm on end user perceived quality. In congested network scenarios, the details of the adaptation algorithm determine the amount of bandwidth consumed by the application as well as the quality of the rendered video stream. HAS will lead to user perceived changes in video quality due to intentional changes in quality video segments, or unintentional perceived quality impairments caused by video decoder artifacts such as pixelation, stutters, or short or long stalls in the rendered video when the playback buffer becomes empty. The HAS adaptation algorithm attempts to find the optimal solution to mitigate the conflict between avoiding buffer stalls and maximizing video quality. In this paper, we present results from a user study that was designed to provide insights into ‘best practice guidelines’ for a HAS adaptation algorithm. Our findings suggest that a buffer-based strategy might provide a better experience under higher network impairment conditions. For the two network scenarios considered, the buffer-based strategy is effective in avoiding stalls but does so at the cost of reduced video quality. However, the buffer-based strategy does yield a lower number of quality switches, as a result of infrequent bitrate adaptations. Participants in buffer-based strategy do notice the drop in video quality causing a decrease in perceived QoE, but the perceived levels of video quality, viewer frustration, and opinions of video clarity and distortion are significantly worse due to artifacts such as stalls in capacity-based. The capacity-based strategy tries to provide the highest video quality possible but produces many more artifacts during playback. The results suggest that player video quality has more of an impact on perceived quality when stalls are infrequent. The study methodology also contributes a unique method for gathering continuous quantitative subjective measure of user perceived quality using a Wii Remote.

CCS Concepts:
- **Networks** → Presentation protocols;
- **Human-centered computing** → Empirical studies in visualization;

Additional Key Words and Phrases: Video Quality, Empirical Evaluation, Dynamic Adaptive Streaming over HTTP, Adaptation Algorithm, Quality of Experience, Streaming Media

ACM Reference Format:

This work has been supported in part by CableLabs, Comcast, and Intel.

Author’s addresses: Ayush Bhargava, James Martin, and Sabarish V. Babu, School of Computing, McAdams Hall, Clemson University, Clemson, South Carolina 29631

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 ACM. 1544-3558/2018/08-ART0 315.00 201: 0000001.0000001
1. INTRODUCTION
As described in Sandvine’s most recent Global Internet Phenomena Report [Sandvine 2018], Internet video streaming, represents 58% of the total downstream volume of traffic on the Internet. Cisco’s Visual Networking Index predicts that video traffic will represent 82% of all IP traffic by 2022 [Index 2016]. Ten years ago the term Internet video streaming assumed UDP transport. Internet video has since evolved considerably from protocols based on UDP, to TCP downloads supporting progressive operation, to a number of approaches based on HTTP adaptive streaming (HAS). The majority of content providers use a standardized version of HAS referred to as Dynamic Adaptive Streaming over HTTP (DASH). DASH provides a standard method to contain, manage, and distribute video content over the Internet [ETSI 2012; Agboma and Liotta 2007]. The standard allows a content provider to create different representations of the content broken up into 'chunks' in a manner that allows any web server to interoperate with a DASH compliant client. The content provider sends a 'Media Presentation Description' or MPD file to the client. This file provides the necessary details to the client including the set of specific web URL’s for the different representations of the content. The MPD file might indicate that portions of the content are located on multiple servers, typically involving a content distribution network (CDN) operator. As we will explain in this paper, the client has a number of design options allowing an implementation to behave differently from other clients. Of particular interest to our research is the choice a content quality adaptation in response to observed network impairment. Our study assumes the simplest system - a single DASH server and a dash client that has a number of experimental design parameter options. For our investigation, we constructed a testbed based on open source components that allowed us to apply different adaptation algorithms in a user study designed to provide specific perceived quality feedback as participants viewed video content streamed in a controlled manner.

A minimal HAS system consists of a server and a client. A HAS server is a web server that holds encoded content packaged in ordered ‘chunks’ referred to as segments. Segments are self-contained such that the client can decode the content independent of previous or next segments. Each segment represents a specific amount of video in terms of viewing time. The size of a segment depends on how it was encoded. A content provider creates multiple representations of the same content, each corresponding to a different level of video quality. The highest quality representation would likely represent high definition video and the lowest quality would likely represent less than standard definition quality. A four second segment (which is a reasonable segment size parameter) encoded at high quality could require an order of magnitude more network bandwidth to stream than that required by the lowest quality representation.

At the start of a session, the client receives a Media Presentation Description (MPD) file for content selection which identifies the possible bitrate options as well as the URL name of all segments. The client consists of a playback buffer, a controller, and a video player. The playback buffer holds a sufficient amount of video data such that if the network temporarily experiences impairment either due to network congestion or connectivity, the player can continue to playback the stream without stalling. The controller monitors the arrival rate of data as well as the state of the playback buffer. It determines when the client should request additional content. The video player consumes video data from the playback buffer at a rate based on the encoded video rate. If the video player requires more data but the playback buffer is empty, the player moves into a stalled state and will not resume rendering video until a configured number of segments have been buffered.

When the network is not congested, the client selects the highest quality video representation. When the network becomes congested, the client could select lower quality video segments. This adaptation algorithm, which operates at the client, makes decisions based on recently observed network and
system conditions. The premise behind HAS is that reducing the requested video quality to match available network bandwidth while minimizing the chances of buffer stalls will lead to improved perceived quality. The HAS application is quite different from an elastic application that does not have any response time or minimum bandwidth requirements. The HAS adaptation algorithm must address the conflict between maintaining high video quality and minimizing buffer stalls. The former goal can be achieved by having the HAS client request the highest quality segments whose anticipated bandwidth requirement matches the predicted available bandwidth for the next segment of time. The latter can be achieved by having the client make adaptation decisions based on the state of the playback buffer. In the literature, these two approaches are referred to as capacity-based and buffer-based adaptation respectively. A capacity-based approach prioritizes high video quality and assumes that the predicted available bandwidth will be sufficient to avoid buffer stalls. Buffer-based adaptation avoids buffer stalls by basing the video quality adaptation on the current state of the playback buffer. The two approaches represent opposite ends of the spectrum of HAS adaptation design strategies. The best approach is likely a compromise that unfortunately is quite difficult to pinpoint. Several reasons for this include 1) each approach is optimal in different conditions; 2) predicting available network bandwidth can be difficult; 3) the content and how it has been encoded impacts the outcome; 4) the best approach needs to map to a precise (and achievable) system optimization; 5) the objectives of an end user might be different from those of an operator.

In the presented research, we explore the human perceived response of video sessions subject to a condition that determines if the capacity-based or buffer-based adaptation is used. We conducted an empirical evaluation involving four carefully designed conditions (organized with a 2 x 2 experimental design). To limit the complexity and scope of the study, we consider only two simple adaptation strategies: a Buffer-Based approach designed to avoid rebuffering stalls at all costs; a Capacity-Based approach designed to adapt based on an estimate of the available bandwidth for the next segment. The former, which we refer to as Strategy 1, is based on the algorithm defined in [Cranley et al. 2006]. The latter, Strategy 2, is based on the algorithm defined in [Stockhammer et al. 2011].

The methodology presents an innovative technique for obtaining continuous perceived quality feedback from participants. To the best of our understanding, the contribution is among the first to compare and contrast the HAS adaptation algorithm with an alternate fundamental approach, and levels of network impairment as conditions in an empirical evaluation measuring user perceived quality of experience (QoE) in a between-subjects manner. Our study goes back to first principles by providing a better understanding of the human response to specific HAS rendering outcomes. The results can perhaps be used to provide data points that a future adaptation algorithm might leverage.

Our research provides insights surrounding the human perception of video rendered by two specific adaptation algorithms by mapping user perceived quality feedback. The broader contribution is the human factors study, which includes an innovative method for incorporating continuous quantitative subjective data collected using a handheld Wii Remote device. All tools and data are publicly available on our project’s web site.

2. RELATED WORK

The study of HAS has been an active area of research in the academic community for many years. The interaction between a HAS client (i.e., the player) and server has been established in recent academic research [Akhshabi et al. 2011; Akhshabi et al. 2012; Alberti et al. 2013; Balachandran et al. 2012]. Measurement studies of commercial HAS systems indicate that the range of bitrates needed to support HD video content is in the range of 0.50 Mbps to 4.8 Mbps [Akhshabi et al. 2011; De Cicco and Mascolo

\[1\text{http://cybertiger.clemson.edu/vss/}\]
More recent work considers 4K video content where bitrates can range up to 15.6 Mbps [Kakhki et al. 2017]. However, as 4K streaming requires Internet access bandwidth of at least 50 Mbps, it is still not widely used. Therefore, we limit our study to video quality up to high definition.

Previous work suggests that a reasonable segment size is between two and ten seconds [Cranley et al. 2006]. Since the client maintains a playback buffer that serves to compensate for the jitter in the traffic arrival process, a playback buffer size ranging from 60 to 240 seconds is reasonable [Cranley et al. 2006; De Cicco et al. 2013]. HTTP-based Adaptive Streaming (HAS) is the dominant Internet video streaming application. One specific HAS approach, Dynamic Adaptive Streaming over HTTP (DASH), is of particular interest as it is standards-based and is widely deployed.

A challenging aspect of HAS is assessing the end user’s perceived quality. The foundations are based on the vast amount of research on video quality assessment that comes primarily from the broadcast video community [De Pessemier et al. 2013; Degrande et al. 2008; Dobrian et al. 2011]. The community has attempted to use objective metrics such as peak signal-to-noise ratio (PSNR), vision-based metrics, packet-stream oriented metrics, pixel based methods, or other stream based methods to automate video quality assessment [Dobrian et al. 2011; Shahid et al. 2014]. The rate of bitrate adaptation is a known factor in the human perceived quality of HAS video [Cranley et al. 2006; Ni et al. 2011]. However, because the metric is typically tied to other system parameters or behaviors it is difficult to find a widely accepted rule of thumb defining a specific maximum rate at which bitrate adaptations lead to reduced perceived quality. Most studies use the rate of bitrate adaptations as a reference to help compare different adaptation algorithms [Bentaleb et al. 2018b]. Some metrics require a frame-by-frame comparison between a reference video and the video under assessment. The lack of a reference video limits the analysis to the observed video, greatly enhancing the flexibility of assessment. However, the challenge with no-reference assessment is being able to accurately differentiate quality degradation from the original content. Subjective video quality assessment involves surveying viewers for their opinion on the quality of the rendered video. While there are some guidelines, subjective video quality assessment can be difficult due to the complexity and difficulties surrounding large scale user studies [Houdaille and Gouache 2012; Huang et al. 2014; Huang et al. 2012].

Given the large scale adoption of HAS systems, there has been an abundance of research that attempts to develop standard QoE metrics that quantify measurable aspects of human perception and then to validate or calibrate the measures based on additional subjective assessments [Alberti et al. 2013; Balachandran et al. 2012; Jiang et al. 2012; Li et al. 2014]. Determining the perceived QoE of a video streaming session is very complex as the assessment depends on many factors including the viewer, the video encoding details and the content [Tavakoli et al. 2014; Rossholm et al. 2014]. The work in [Seufert et al. 2014] provides a recent, comprehensive survey on the topic.

The literature reflects a set of accepted ‘best practices’ to HAS system developers:

—The percentage of time spent buffering has the largest impact on user engagement across all types of content [Huynh-Thu and Ghanbari 2006].

—Users are more tolerant of reduced video quality that is stable than a video session that frequently changes between various levels of quality ranging from reduced video to high quality video [Huysegems et al. 2012; ITU-T RECOMMENDATION 1999; Li et al. 2014; Liu et al. 2012; Sector 1998].

—Users prefer a single long stall rather than multiple short stalls [Jackson et al. 2015].

HAS represents an application with variable bandwidth demands and that operates in a network subject to control by TCP and other network management components over the path. Incremental studies that consider various aspects of the system design continue to be published [Timmerer et al. 2014].
Comparative Evaluation of User Perceived Quality Assessment of Design Strategies for HTTP-Based Adaptive Streaming

2014; Sani et al. 2017; Bentaleb et al. 2018a; Alzahrani et al. 2018; Dubin et al. 2019]. However, being based on TCP, HAS suffers from performance issues like the number of RTTs for the handshake, the Head of Line Blocking, etc. [Cook et al. 2017]. Therefore, changes to TCP/IP have been proposed to address these issues [Cook et al. 2017; Ameur et al. 2017]. An alternative transport protocol referred to as QUIC has been proposed [Langley et al. 2017; Mirko Palmer 2018]. QUIC addresses problems caused by TCP, however, it is not clear if QUIC improves a HAS system [Kakhki et al. 2017; Bhat et al. 2017]. QUIC does not address in anyway the core issue of how to dynamically select the quality of video content.

In summary, the HAS adaptation algorithm is an open issue. Prior research has led to a set of loosely defined engineering guidelines for HAS adaptation algorithms. For example, the work in [Huynh-Thu and Ghanbari 2006] suggests that, “the percentage of time spent buffer (buffering ratio) has the largest impact on user engagement.” This clearly motivates the proposed strategy that ties the adaptation directly to the state of the playback buffer. However, the adaptation must also consider (maximize) the average content quality as well as minimize excessive quality changes. Finding an adaptation algorithm that achieves acceptable perceived quality based on multiple objectives is an open research issue. The research presented in this paper presents a novel user study method designed to provide more precise guidelines for the HAS adaptation.

3. SYSTEM DESCRIPTION

The system under study models a user viewing video over the Internet. An example could be a broadband Internet access subscriber viewing Netflix content. In our study, participants were subjected to a controlled viewing experience. Throughout the experience, we obtain quantitative and qualitative information that collectively provides feedback on the participant’s perceived quality. We selected a short sci-fi movie with a run time of 12 minutes and 15 seconds for the study. The film, 'Tears of Steel' is available under the terms of Creative Commons. We wanted to use a clip that was rich in content, had lots of visual effects and frequent changes with respect to on-screen objects. Using such a clip ensured that even the slightest degradation could be identified by the viewer and the playback tested the limits of the design strategy on the viewer’s perception and reaction.

We obtained a high quality (4K quality) uncompressed version of the content. We used the open source ffmpeg and webM tools to create H.264 encoded, DASH compliant segment files. The representations ranged from 0.5 Mbps to 4.8 Mbps. The highest quality representation encoded at a bit rate of 4.8 Mbps represents roughly the same level as High Definition (HD) quality video. We set the segment size to 4 seconds and the maximum playback buffer size to 120 seconds. These configuration settings are based on recent measurement studies of deployed HAS systems engineered to deliver high definition streams [De Cicco et al. 2013; Akhshabi et al. 2011; De Cicco and Mascolo 2014]. The representations were viewed using a modified version of an open source JavaScript HAS player, extended the player with implementations of the Buffer-Based Adaptation (i.e., design strategy 1) and the Capacity-Based Adaptation (design strategy 2) based on [Cranley et al. 2006; Stockhammer et al. 2011] respectively.

---

1Information from the film developer is available at https://mango.blender.org/
2The dash.js open source project is available at https://github.com/Dash-Industry-Forum
3Further information about the algorithms and the player implementations can be found at our project web site located at http://cybertiger.clemson.edu/vss
3.1 Experimental Setup
The viewing experiences were conducted in a small theater-like setup in the Digital Production Arts facility at Clemson University. The facility offers a large screen akin to a home theater viewing system, with high-end projection. In recent years, as displays have become cheaper, the consumer market for displays has seen a shift from small single screen displays to multi-screen bezel less displays scaling over 200 inches, large-scale 3D capable projectors and curved displays scaling over 105 inches. Many users have also started watching content in a large theater like virtual environment utilizing Head-Mounted Displays like the Oculus Rift and the HTC Vive. Thus, we believe our setup evaluates user preferences of viewing content in current and future home entertainment viewing scenarios. The participants were asked to sit in the center of the room, viewing the content from an optimal vantage point at about 6 meters away from the screen. They were at a comfortable distance, where a full view of the screen could be maintained.

To obtain real-time continuous subjective quantitative assessment from participants, we used a continuous feedback method involving a Wii Remote device (referred to as a WiiMote). During a viewing experience, participants were asked to indicate their perceived levels of poor quality of the video by tilting the hand held WiiMote to the right or left (further details in section 3.3). Participants were not allowed to pause the video. Before the video was played, the participant was asked to fill out surveys and instructed on how to hold and use the WiiMote during the video playback.

3.2 Video Conditions
As mentioned earlier, an actual HAS system involves client players interacting with content servers located in the internet. For our study, we needed to reproduce aspects of the actual system but in a manner that supports a repeatable experimentation platform with sufficient control. Our process involved two steps: 1) Platform emulation and calibration; 2) Production of the video conditions. Following these steps, we produced four specific videos that represent the user study conditions that form the basis of the 2x2 experimental study design.

3.2.1 Platform Emulation and Calibration. The experimental setup used to emulate realistic HAS Streaming includes a Content Server that holds the ‘Tears of Steel’ content in DASH formats. A HAS client that runs our modified adaptation algorithms (Windows PC) and a Linux PC configured as a router using netem. All physical networks were connected via gigabit ethernet. We used netem to add controlled levels of artificial packet loss and latency to network traffic. Applying different loss patterns is an effective method for emulating changes in the available bandwidth over the path that is available to the HAS streaming session. Setting netem to use the Bernoulli loss model configured with a 3% loss rate leads to streaming session with a small level of observable artifacts, low impairment dynamic. Increasing the netem average loss rate to 9% corresponds to a significantly large number of artifacts, high impairment dynamic. The resulting variation in available TCP bandwidth for the low impairment conditions at 3% packet loss and the high impairment conditions at 9% packet loss reflected realistic streaming scenarios such as those described in [Akhshabi et al. 2012; Müller et al. 2012].

3.2.2 Production of the video conditions. Each of the four video conditions (identified in Table I) represents a specific sample path (determined by the emulation and calibration step) of the random process that defines HAS streaming sessions. The methodology could have had each participant’s streaming session experience based on an actual streaming session. However, it is difficult to have each session, even those configured for the same conditions, reproduce identical random outcomes.

---

6 The DPA Theater has a Christie D4K256 projector operating in 2K video mode
7 https://wiki.linuxfoundation.org/networking/netem

ACM Transactions on Applied Perception, Vol. 0, No. 0, Article 0, Publication date: August 2018.
Table I. User Study Conditions - 2 x 2 Experiment design.

<table>
<thead>
<tr>
<th>Streaming Strategy /Impairment</th>
<th>Strategy 1 (Buffer based)</th>
<th>Strategy 2 (Capacity based)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>S1-HI</td>
<td>S2-HI</td>
</tr>
<tr>
<td>Low</td>
<td>S1-LO</td>
<td>S2-LO</td>
</tr>
</tbody>
</table>

Table II. Video Experience Definition and Summary

<table>
<thead>
<tr>
<th>Quantitative Measure</th>
<th>S1-HI</th>
<th>S2-HI</th>
<th>S1-LO</th>
<th>S2-LO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average playback rate (Mbps)</td>
<td>1.024</td>
<td>1.483</td>
<td>2.988</td>
<td>3.401</td>
</tr>
<tr>
<td>Average playback buffer size (seconds)</td>
<td>64.11</td>
<td>18.40</td>
<td>56.4</td>
<td>8.87</td>
</tr>
<tr>
<td>Number of quality changes</td>
<td>9</td>
<td>21</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Number of long stalls</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Number of short stalls</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Number of frame drops</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Total number of artifacts</td>
<td>13</td>
<td>43</td>
<td>2</td>
<td>29</td>
</tr>
</tbody>
</table>

To ensure that the video experiences were delivered to participants in a repeatable manner, a video recording was created for each of the four possible video conditions. The recordings were created using the SSR screen capture software in a high quality AVI video file format. These files were placed on a machine that was directly connected to the projector. Since we played back a recorded video of the rendering, we modified it to remove the startup buffering time purposely so it was not a factor in the investigation.

We used the following measures to quantify the rendered quality of each video condition. The first three measures were obtained directly from the HAS client. We instrumented the client to maintain the measures and to periodically log samples. We have a visualization of each video condition that includes the measure samples at http://cybertiger.clemson.edu/vss/UserStudy.html. The other measures were artifacts that we assessed with a group of viewers (not from the study participant group).

—**Average playback rate**: Average rate (in bps) at which the video was played by the player. If the video player stalls, samples would not be recorded (i.e., rebuffering events will not impact the measure).

—**Average playback buffer size**: Average length of video (in seconds) available in the player’s memory/buffer that is ready to be played.

—**Number of quality changes**: Total number of resolution upgrades or downgrades in the video.

—**Number of rebuffering events**: classified as a ‘long stall’ (of 5 seconds or longer) or a ‘short stall’ (lasting less than 5 seconds)

—**Number of frame drops**: Total number of times a frame was dropped or skipped by the player.

We examined each of the four video conditions and each observed artifact was categorized into one of three assessed artifacts. Since each video had a varying number of artifacts that were subject to interpretation, we had three team members perform this activity separately followed by a team discussion to ensure we produced an accurate list of artifacts. Table II summarizes the quantitative characteristics from each of the four videos using the framework provided by the measures. The results are very much inline with the published results of measurement studies of commercial HAS systems [Akhshabi et al. 2012; Müller et al. 2012; Degrande et al. 2008].

3.3 WiiMote Setup

The WiiMote device provides an effective, unobtrusive method for recording continuous participant responses to the stimuli via a gesture. An application, developed using the Unity game engine, was created to record the responses. The recorded versions of the four video experiences can be viewed at http://cybertiger.clemson.edu/vss/

The Simple Screen Recorder is available at http://www.maartenbaert.be/simplescreenrecorder/

Unity: http://unity3d.com/
used to poll the WiiMote and record the participants’ continuous hand gestural feedback of the perceived quality of the video viewing experience into a log file11. The WiiMote communicates with the application using a Bluetooth connection. To get continuous quality feedback from participants, they were instructed to hold the device in an upright position and rotate it from left to right (i.e. from an angle of 0-good quality & low frustration to 180-poor quality & high frustration). The WiiMote’s rotation about the axis along the participant’s wrist was recorded every 50 milliseconds. All participants received some training prior to the start of the experiment in order to gain familiarity with using the WiiMote device in providing continuous feedback. A time stamped record of the participants’ response was logged. The starting time of the video session was recorded, allowing us to interpret the WiiMote data as a fixed increment time series.

4. STUDY DESIGN

4.1 Hypotheses

The study is designed to formulate guidelines as to how a HAS adaptation algorithm should address the conflicting requirement of maximizing video player quality and minimizing the frequency of buffer stalls. We identify the following hypotheses:

— **Hypothesis 1**: In high impairment network conditions, perceived quality is greater when using Strategy 1 (Buffer based) than Strategy 2 (Capacity based).

— **Hypothesis 2**: In low impairment network conditions, perceived quality is greater when using Strategy 2 (Capacity based) than Strategy 1 (Buffer based).

— **Hypothesis 3**: Users become less tolerant of lower quality overall as the streaming session progresses.

— **Hypothesis 4**: For both network conditions (high or low), users respond more readily to buffer stalls as compared to other artifacts.

4.2 Participants

A within-subjects design was used for this study. Using two levels of network impairment and two streaming algorithms, we conducted a 2 x 2 factorial design experiment based on four conditions that reflect combinations of network impairment and the HAS adaptation strategy. A third variable of viewing time was also used to break down the video for better measurement of viewer responses. Apart from the post-survey QoE assessment and the continuous feedback, we also collected demographic and behavioral data, personality data and current mood state data.

A total of 33 participants were recruited, 31 males and 2 females. They included a mix of graduate and undergraduate students from Clemson University. The participant ages ranged from 18 years to 30 years with a mean of 22 years. Each participant was expected to attend 2 sessions on separate days, each lasting approximately 30 to 40 minutes. The second session was conducted at least two days after the participant’s first session to eliminate any carryover effects from the previous session. No participant saw the same video twice. This resulted in a total of 66 video sessions, 59 of them contributed data to the analysis. Data from 7 sessions was excluded from the analysis due to unexpected system interrupts and desynchronization of sound during video playback. The sample data distribution of sessions between conditions was near equal. S1-HI, S2-HI and S2-LO had 15 sessions each and S1-LO had 14 sessions.

4.3 Methodology

Participants were required to attend two sessions. The procedure for the first session was as follows:

---

11Information about the Wiimote is available at http://wiibrew.org/wiki/Wiimote
(1) The participant was handed an informed consent describing the purpose of the study.
(2) The participant then filled out 3 pre-surveys pertaining to their personality, current mood state, demographics and general streaming video content watching habits.
(3) Following this, instructions were given on how to use the WiiMote while watching the video. The participants were simply asked to watch the video and convey their perceived QoE using the WiiMote. They were not given any instructions on what artifacts to look for or if they needed to pay attention to the plot.
(4) Next, a video from one of the 4 conditions was randomly selected to be presented to the participant.
(5) After the video, participants filled out a post-survey pertaining to his/her QoE.

The second session was conducted at least two days after the participant’s first session in order to eliminate any carryover effects from the previous session. The procedure for the second session was as follows:

(1) The participant only filled out the current mood state pre-survey.
(2) Then one of the 3 remaining videos was presented with the same instructions in a manner similar to the first video viewing session.
(3) The participant then filled out the post-survey.
(4) Finally, the participant was debriefed regarding the purpose of the experiment.

4.4 Measures

4.4.1 Quantitative Measures.

4.4.1.1 Subjective. The post-survey was used to collect quantitative data about the participant’s QoE. We made use of a 5-point Likert scale to help participants represent frustration (feeling of being annoyed with the video playback experience), distortion (presence of artifacts that would degrade or distort the video playback), and video quality (how clear or crisp the video is. Higher quality implies higher resolution) by choosing an option that best aligned with their view.

To make it simpler for the participants to comprehend the scale, the values were replaced with phrases. For frustration the scale read extremely frustrated (1), somewhat frustrated, neutral, somewhat satisfied and extremely satisfied (5). Higher the frustration, smaller the value it represented on the Likert scale. This was also true for distortion with phrases extremely distorted (1), somewhat distorted, neutral, somewhat clear and extremely clear (5). The inverse was true for video clarity. Higher the clarity, higher the value on the Likert scale. The phrases read extremely blurry (1), somewhat blurry, neutral, somewhat blurry and extremely blurry (5) The details of analysis on this data can be seen in the results section (5.1.1).

4.4.1.2 Continuous Subjective. Multiple post-processing steps were applied to each participant’s continuous feedback data provided using the handheld WiiMote device to filter out the noise and normalize scores. Post normalization, a custom automatic response detection program was used to identify artifacts in the video responded to by participants.

To filter and smooth the raw data for each participant, a 1 second sliding temporal window filter similar to the moving average technique was applied. Iterating through each sample (skipping samples in the first 0.5 seconds and the last 0.5 seconds of the video), an average is calculated for all samples within a range of +/- 0.5 seconds of the current sample time. Equation 1 represents the calculation for each sample at time $t$, $V_t$ represents the smoothed value of the perceived quality at $t$. A value approaching 0 represents good quality and a value approaching 180 represents bad quality.
\[ V_t = \frac{\sum_{t-0.5 \leq t < t+0.5} v_t}{\text{Time stamps recorded between } t-0.5 \text{ & } t+0.5} \] (1)

The data was normalized to make the range of recorded WiiMote samples consistent between participants for analysis. We analyzed the continuous data in two different ways. First, we divided each video into 3 equal time intervals, beginning, middle and end, approximately 4 minutes each. The values from each interval were then averaged for each participant and analyzed. For a participant P, using \( d \) to represent the duration of video playback and \( v_t \) to represent the smoothed perceived quality measured at time \( t \), we define three average measures of perceived quality based on three time segments:

\[
\begin{align*}
\text{Avg}_{\text{beg}} &= \frac{\sum_{0 \leq t < d/3} v_t}{\text{Time stamps recorded between } 0 \text{ to } d/3} \\
\text{Avg}_{\text{mid}} &= \frac{\sum_{d/3 \leq t < 2d/3} v_t}{\text{Time stamps recorded between } d/3 \text{ to } 2d/3} \\
\text{Avg}_{\text{end}} &= \frac{\sum_{2d/3 \leq t < d} v_t}{\text{Time stamps recorded between } 2d/3 \text{ to } d}
\end{align*}
\] (2-4)

where,

\( \text{Avg}_{\text{beg}} \) is the average perceived level of poor quality for approximately the first 4 minutes of the video.

\( \text{Avg}_{\text{mid}} \) is the average perceived level of poor quality for approximately the middle 4 minutes of the video.

\( \text{Avg}_{\text{end}} \) is the average perceived level of poor quality for approximately the last 4 minutes of the video.

Secondly, we calculated the total number of artifacts a participant responded to during video playback. This was done using an automated program. The program checks for maxima and minima once an artifact has occurred. If a maximum or minimum did occur and met a specified threshold, it was considered as a reaction to the corresponding artifact. The heuristics for response detection were based on observations made from the normalized graphs. Table III describes the observations and the heuristic associated.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Heuristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants tend to move towards local maximum for artifacts that degrade the QoE.</td>
<td>A local maximum was considered for stalls, frame drops and quality degrades as these degraded QoE.</td>
</tr>
<tr>
<td>Participants tend to move towards local minimum for artifacts that improve the QoE.</td>
<td>A local minimum was considered for quality upgrades.</td>
</tr>
<tr>
<td>Participants take about 1-2 seconds to start reacting to an artifact and it takes around 4-7 seconds to see a maximum or minimum.</td>
<td>The time span considered for a local maximum or minimum was 7 seconds from the start time of the artifact.</td>
</tr>
<tr>
<td>The magnitude of response at the start of the video is not very large. Multiple artifacts that occur within a span of 5 seconds or less might have a combined response of high magnitude.</td>
<td>The threshold for maxima and minima being considered as responses was set to 0.05 units on a normalized scale. The value of the maximum or minimum was compared to the value at the start of the artifact for detection.</td>
</tr>
</tbody>
</table>
The process to quantify the responses to the artifact using the program mentioned above gave us the number of artifacts a participant reacted to during the video viewing experience. These responses were marked on the graph using colored spherical marker tags placed right above the maximum or minimum corresponding to that artifact (see Figure 1).

Different colored and sized marker tags were used to represent responses to different artifacts. The color of tags correspond to the color of the artifact whose response they represent. Thus, the red tags represent responses corresponding to stalls, the orange tags correspond to frame drops, the yellow tags correspond to quality degrades and the green tags correspond to quality upgrades. Different sizes were used to make it easier to read responses when the same peak was considered for 2 different artifacts in case both the artifacts occurred within a span of 5 seconds or less. In a separate validation process, the automatic method of identifying local maximum and minimum were verified for correctness by applying the automatic detection and tagging method on experimenter created data, and was found to be robust in detecting instances of perceived poor quality or improvement in perceived quality corresponding to the artifacts found in the videos.

A summary of each condition can be seen in Table IV (section 5.2). Since, the actual number of artifacts present in the video was different for each condition, the participant responses were converted into percentiles. This was done by dividing the number of artifacts participants responded to by the total number of artifacts present in the video he/she viewed. We used the percentage of total artifacts responded to by the participant for studying the effects of streaming strategy on the QoE. The results of the analysis are explained in detail in section 6.

4.4.2 Qualitative Measure. The post-survey also asked participants for comments on their experience towards the beginning of the video, the end of the video and overall. Comments related to the
video quality and the artifacts that occurred during the video were selected from each video condition for further analysis. A detailed analysis of the qualitative results can be found in section 7.

5. RESULTS

5.1 Quantitative Subjective Results

Effects of Network Algorithm, Impairment, Session and sampling Time on user’s Frustration, Clarity, Distortion, and Perceived Levels of Quality of video experience

For each of the dependent measures (frustration, clarity and distortion), we performed a 2 x 2 x 2 repeated measures Analysis of Variance (ANOVA) statistical analysis on the beginning and end scores, and a 2 x 2 repeated measures ANOVA statistical analysis on the overall scores, after carefully verifying the suitability of the analysis methods by consulting an experimental statistics expert. Then, we verified that the underlying assumptions for parametric ANOVA analysis were met, namely descriptive statistics revealed that the samples were normally distributed and error variance in groups of sample data were equivalent. We ensured that Box’s test of equality of covariance matrix was not significant, and conducted Levene’s test of equality of variance on the sample data to ensure that there was near equal variance in the samples prior to conducting parametric analysis on the data. The within-subjects repeated measures factors were sampling time (beginning and end), algorithm namely buffer-based adaptation Strategy 1 (S1) and capacity-based adaptation Strategy 2 (S2), as well as impairment (high and low). Greenhouse-Geiser adjusted degrees of freedom were considered when Mauchly’s test of sphericity was found to be significant. Post-hoc analysis on pairwise comparisons were conducted using Bonferroni-adjusted type I error method.

5.1.1 Frustration. The ANOVA analysis on the beginning and end scores did not reveal a significant main effect of time, but revealed a significant sampling time by algorithm interaction on frustration QoE scores $F(1, 55) = 4.105, p = .048, \eta^2 = .07$, see Figure 2 left. The data did not reveal any other
significant main or interaction effects. In order to further examine the interaction effect, pairwise comparison with Bonferroni method was conducted. The analysis revealed that frustration scores towards the end of the video were significantly higher for Strategy 1 (M = 3.24, SD = 1.29) than Strategy 2 (M = 2.55, SD = 1.15), \( p < .05 \).

The ANOVA analysis on the overall scores revealed a significant main effect of impairment \( F(1, 55) = 15.234, p < .01, \eta^2 = .22 \) and a significant interaction effect of algorithm by impairment \( F(1, 55) = 6.633, p < .05, \eta^2 = .11 \), see Figure 2 right. A post-hoc comparison revealed that mean frustration scores overall within the low impairment condition were significantly higher for Strategy 1 (M = 4.14, SD = 1.03) than Strategy 2 (M = 3.0, SD = 1.15), \( p < .05 \). The overall scores within the Strategy 1 condition were significantly higher for low impairment (M = 4.14, SD = 1.03) than high impairment (M = 2.4, SD = .91), \( p < .01 \). Recall that low frustration QoE scores indicate high levels of viewer frustration and high frustration QoE scores indicate low levels of viewer frustration in the video viewing experience.

5.1.2 Video Clarity. The ANOVA analysis on the beginning and end scores revealed a significant main effect of time, \( F(1, 49) = 20.478, p < .01, \eta^2 = .295 \). The analysis also revealed a significant interaction effect of sampling time by impairment \( F(1, 49) = 5.491, p = .023, \eta^2 = .1, \) see Figure 3 left. In order to further examine the interaction effect, pairwise comparison using Bonferroni method was conducted. The analysis revealed that the mean video clarity scores in the end were significantly higher for the low impairment condition (M = 3.67, SD = 1.2) than the high impairment condition (M = 2.63, SD = 1.09), \( p < .05 \). The mean video clarity scores within the low impairment condition were significantly higher towards the end (M = 3.72, SD = 1.19) than at the beginning (M = 2.38, SD = 1.32), \( p < .01 \).

The ANOVA analysis on the overall scores revealed a significant main effect of impairment \( F(1, 49) = 21.782, p < .01, \eta^2 = .31 \) and a significant interaction effect of algorithm by impairment \( F(1, 49) = 6.633, p = .023, \eta^2 = .1, \) see Figure 3 right. Post-hoc analysis revealed that, for Strategy 1, the overall scores were significantly higher under low impairment (M = 4.0, SD = 1.31) than in the high impairment condition (M = 2.25, SD = .87), \( p < .05 \). Also, the overall clarity scores in the high impairment condition
Fig. 4. Interaction graph showing mean distortion QoE scores in the beginning and end in the high and low network impairment conditions

were significantly higher for Strategy 2 (M = 3.18, SD = 1.08) than Strategy 1 (M = 2.15, SD = .89), p < .05.

5.1.3 Distortion. The ANOVA analysis on the beginning and end scores did not reveal a significant main effect of time or session, but revealed a significant sampling time by impairment interaction on distortion QoE scores $F(1, 54) = 7.937, p = .007, \eta^2 = .13$, see Figure 4. In order to further examine the interaction effect, pairwise comparison using the Bonferroni method was conducted. The analysis revealed that the mean distortion scores towards the end were significantly higher for the low impairment condition (M = 3.59, SD = 1.38) than the high impairment condition (M = 2.48, SD = 1.02), p = .002. The mean distortion scores within the low impairment condition were significantly higher towards the end (M = 3.59, SD = 1.38) than at the beginning (M = 2.62, SD = 1.18), p = .008. Recall that low distortion QoE scores indicate high levels of distortion and high distortion QoE scores indicate low levels of distortion in the video.

The ANOVA analysis on the overall scores revealed a significant main effect of impairment $F(1, 54) = 35.026, p < .01, \eta^2 = .34$.

5.2 Quantitative Continuous Results

5.2.1 User perceived QoE recorded using Wii Remote. After the participants’ continuous response was filtered and normalized based on the technique highlighted in section 4.4.1.2, the following data analysis was performed. Final scores were between 0: high video quality (low levels of viewer frustration) to 1: low video quality (high levels of viewer frustration). We examined to what extent users perceived the quality in the beginning (first four minutes), middle (middle four minutes), and end (last four minutes) of the video viewing experience overall as a function of the impairment (high versus low) and algorithm (S2 versus S1) on the filtered and normalized quality rating (0 high quality/low frustration 1 low quality/high frustration). Continuous measure of user quality was averaged across the four minute periods in the beginning, middle and end time periods of the video viewing experience.

The average perceived quality measures were treated with a three way $2 \times 2 \times 3$ repeated measures ANOVA, with impairment (2 levels high and low), algorithm (2 levels S1 and S2) and time of video viewing (3 levels - beginning, middle, end) as the independent factors. Greenhouse-Geiser adjusted degrees of freedom were considered when Mauchly’s test of sphericity was found to be significant. Pairwise comparisons in the post-hoc analysis with Bonferroni-adjusted type-I error were computed to verify significant pairwise effects. Recall that low continuous measure scores reveal high video quality or low levels of frustration, and high scores reveal poor video quality or high levels of frustration.
The ANOVA analysis revealed a significant main effect of time, $F(2, 108) = 20.13, p < .001, \eta^2 = .27$. The main effect of algorithm and impairment were not significant. However, sampling time by algorithm interaction revealed a significant effect $F(2, 108) = 5.83 p = .004, \eta^2 = .09$ (Figure 5 left), and sampling time by impairment interaction was also significant $F(2, 108) = 5.79 p = .004, \eta^2 = .09$ (Figure 5 right). Also, the three way interaction of sampling time by algorithm by impairment revealed a significant effect $F(2, 108) = 3.42, p = .036, \eta^2 = .06$. In order to further examine the interaction effects, pairwise comparison with Bonferroni method was conducted.

Pairwise comparisons conducted to explore the time by algorithm interaction revealed that, for Strategy 1, the scores towards the beginning of the video (M = .38, SD = .19) were significantly lower than the scores towards the middle of the video (M = .45, SD = .24), $p < .05$. For Strategy 2, the scores towards the beginning of the video (M = .30, SD = .13) were significantly lower than both the scores towards the middle of the video (M = .45, SD = .15), $p < .01$, and the scores towards the end of the video (M = .46, SD = .16), $p < .01$.

Pairwise comparisons conducted to explore the time by impairment interaction revealed that the scores towards the middle of the video were significantly higher for the high impairment condition (M = .49, SD = .17) than those for the low impairment condition (M = .36, SD = .21), $p < .05$. Within the high impairment condition, the scores towards the beginning of the video (M = .38, SD = .16) were significantly lower than both the scores towards the middle of the video (M = .49, SD = .17), $p < .01$, and the scores towards the end of the video (M = .46, SD = .17), $p < .05$. Also, within the low impairment condition, the scores towards the end of the video (M = .45, SD = .23) were significantly higher than both the scores towards the beginning of the video (M = .30, SD = .17), $p < .01$, and the scores towards the middle of the video (M = .36, SD = .21), $p < .05$.

Pairwise comparisons within the levels of impairment or algorithm revealed that, for Strategy 2 under high impairment, the scores towards the beginning of the video (M = .35, SD = .13) were significantly lower than both the scores towards the middle of the video (M = .54, SD = .12), $p < .01$, and the scores towards the end of the video (M = .44, SD = .12), $p < .05$. Also, the scores towards the middle of
Table IV. Artifacts and response summary for each condition

<table>
<thead>
<tr>
<th>Artifacts &amp; Response Summary/ Condition</th>
<th>S1-HI</th>
<th>S2-HI</th>
<th>S1-LO</th>
<th>S2-LO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Stalls</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Number of Frame Drops</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Number of Quality Degrades</td>
<td>6</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of Quality Upgrades</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>41</td>
<td>2</td>
<td>29</td>
</tr>
</tbody>
</table>

Mean proportion of times participants reacted to artifacts: 31.32% \( \pm \) 21.43% for S1-HI, 51.87% \( \pm \) 15.77% for S2-HI, 52.57% \( \pm \) 41.42% for S1-LO, 56.09% \( \pm \) 24.68% for S2-LO.

the video \( (M = .54, SD = .12) \) were significantly higher than the scores towards the end of the video \( (M = .44, SD = .12), p < .05 \). A similar trend was observed for Strategy 2 under low impairment. The analysis revealed that the scores towards the beginning of the video \( (M = .26, SD = .12) \) were significantly lower than both the scores towards the middle of the video \( (M = .36, SD = .13), p < .05 \), and the scores towards the end of the video \( (M = .48, SD = .19), p < .01 \). The scores towards the middle of the video \( (M = .36, SD = .13) \) were also significantly lower than the scores towards the end of the video \( (M = .48, SD = .19), p < .05 \).

5.2.2 Automatic classification of user reaction to artifacts from Wii Remote data. To analyze the quantitative results obtained from the automatic classification of user’s reactions to the presence of the artifacts in the continuous user response through the handheld input device, we performed a block frequency analysis on the total response percentile in each condition. The means and standard deviation for each condition are given in the table above, Table IV.

The test showed that there was a statistically significant difference in response percentiles between the different video streaming algorithms under high network impairment, \( \chi^2 (2) = 5.572, p = .018, \eta^2 = .19 \), with a mean rank response percentile of 11.14 for Strategy 1 and 18.60 for Strategy 2. The video streaming strategy accounts for 19% of the variance in the response percentiles. Participants seem to react more frequently to artifacts while viewing videos streamed using Strategy 2 under high network impairment resulting in higher overall frustration levels.

6. QUALITATIVE RESULTS

The comments from the post-survey were used to qualitatively assess the perceived level of quality of the overall experience.

For the first streaming condition i.e. Strategy 1 under high network impairment (S1-Hi), some viewers thought that a smooth streaming video is better than having stalls and that the overall audio and video quality were great; “Smooth is better than stall”, “The video overall looked and sounded great”. While other viewers did not like the video quality as much and would have switched to a better source; “If I was streaming this, I would have tried to find another source and only streamed from here if it was the only option”, “video quality seemed much poorer this time around overall”, “Quality could be better”.

For the second condition i.e. Strategy 2 under high network impairment (S2-Hi), comments were critical of the frequent stalls and the low video quality. Viewers felt that the magnitude of impairment was enough to prevent them from understanding the plot of the movie; “pauses in video were most frustrating”, “Well, it just looks grainy. Especially on such a large screen”, “Between the buffering and
grainy clarity, I was not able to enjoy the movie”, “Pauses in video were more prominent by the end, and it detracted from my following the plot of the movie”.

Viewers from the third condition, Strategy 1 under low network impairment (S1-Lo), thought that the overall experience was extremely great and nothing was frustrating overall; “Extremely great video quality and audio quality”, “I could tell what was going on and it was a good short film. Nothing seemed too frustrating overall”, “Even though the beginning was blurry, an overwhelming majority of the film was very crisp”. But, few viewers thought that the sound quality was random and the video had a lot of distortion; “Audio volume seemed random through the whole thing. Lots of tearing and distortion in the video”.

For the last condition, Strategy 2 under low network impairment, viewers seem to prefer the video quality but were frustrated with the frequent stalls and frame drops; “Overall, I was impressed with the quality. Other than the beginning quality, and ending lagging/chopping”, “When the video started to freeze and skip it became much more difficult to follow along with what was happening on screen”; “The video image was very good, but the lagging/chopping was very frustrating”, “Towards the end of the video, the picture and audio seemed to freeze at certain moments and then skip ahead, which made viewing and understanding what was happening a little more difficult”.

Ranking the four video conditions based on viewer comments, we observe that, S2-Hi seems to provide the worst viewing experience. S1-Hi is better than S2-Hi with smoother video playback. S1-Lo is better than S1-Hi with better video quality and S2-Lo provides the best playback of the four with an exceptionally great overall experience.

7. DISCUSSION

From the Quantitative Subjective Results section (5.1.1), we observe that the frustration QoE scores towards the end of the video in Strategy 2 were lower than in Strategy 1. From Quantitative subjective measures (4.4.1), higher frustration represents lower scores on the QoE Likert scale, thus the above observation supports the first hypothesis which states that the perceived quality is greater for Strategy 1 as compared to Strategy 2. The first hypothesis is also supported based on the quantitative continuous results (section 5.2.2) i.e. in the high network impairment condition, participants responded to artifacts more frequently while viewing videos streamed using Strategy 2 resulting in higher frustration levels in Strategy 2 (capacity based) as compared to Strategy 1 (buffer based). The participant comments in the qualitative results (section 6) also show a similar trend with viewers favoring videos streamed under high impairment using Strategy 1 rather than Strategy 2, hence supporting the first hypothesis. A possible explanation for this trend could be the presence of a larger number of artifacts in videos streamed using Strategy 2. Videos streamed using Strategy 2 also tend to have more buffering stalls among other artifacts which have been reported to have the largest impact on user engagement [Huynh-Thu and Ghanbari 2006].

The overall frustration scores (section 5.1.1) observed under low impairment condition are lower for Strategy 2 than Strategy 1, hence not supporting the second hypothesis which states that in the low impairment condition, the perceived quality is less negatively impacted by Strategy 2. However, participant comments (section 6) support the second hypothesis. Although the video streamed using Strategy 2 under low impairment exhibits more artifacts than Strategy 1 (section 5.2.2, Table IV), viewer comments suggest that the quality was better for Strategy 2 streamed videos in comparison to Strategy 1. This makes sense due to higher observed playerRate in Strategy 2. Since Strategy 2 in low impairment situations tries to jump to the highest available bandwidth, the playback is of higher quality.

In the low network impairment condition, the distortion scores observed in the beginning of the video are lower than those observed at the end (section 5.1.3). For distortion, the higher the values on the
Likert scale, the lower the distortion (section 4.4.1), thus portraying a higher level of perceived quality as the video progressed. This does not support the third hypothesis which states that the perceived level of poor quality is expected to increase with the number of artifacts overtime. Same is true for the clarity scores obtained (section 5.1.2), with scores being higher towards the end of the video as compared to the beginning. The third hypothesis is not supported as higher clarity scores represent a higher value on the Likert scale. One possible explanation for these results could be that when measured after the viewing experience, participants are more inclined to remember the effects of the artifacts in the beginning of the viewing experience more so than towards the end. This could be due to the novelty factor that facilitates better recall of the video viewing experience in the beginning, and through acclimation over time.

Although the quantitative subjective results are not supportive of the third hypothesis, the quantitative continuous results (section 5.2.1) and qualitative results (section 6) do support it. An analysis of the continuous measure of users' perceived levels of poor quality, higher levels representing higher frustration, revealed that the perceived levels of poor quality for Strategy 1 and Strategy 2 were lower in the beginning as compared to the middle and the end under both high and low impairment conditions. In the high impairment condition, perceived levels of poor quality was low in the beginning, but increased towards the middle and end of the video viewing experience. Since the recorded response of the participants is immediate rather than after-the-fact and is less prone to short term memory effects, we believe that the continuous measure is more robust than the subjective post-experience questionnaire. Multiple user comments from different viewing conditions convey that stalls became more frequent towards the end along with “chopping” and other sound related issues which made it difficult for the users to understand the plot of the movie.

The analysis of the automatic classification of user responses from quantitative continuous results (section 5.2.2) yielded that, under high network impairment, participants responded to artifacts more frequently in Strategy 2 as compared to Strategy 1 implying higher perceived levels of poor quality for Strategy 2. This is partially supportive of the fourth hypothesis, which states that users respond to buffer stalls more readily as compared to other artifacts in both network conditions, as S2-Hi condition has considerably more buffer stalls than S1-Hi. Overall, when presented with high impairment situations, perhaps arising due to network congestion, the quantitative and qualitative results of our user study suggests that participants seem to favor the buffer based Strategy 1 over the capacity based Strategy 2 for video streaming.

8. CONCLUSION AND FUTURE WORK

We conducted a within-subjects user study involving a 2 x 2 factorial methodology based on the level of network impairment (high impairment or low impairment) and the choice of adaptation design strategy (buffer-based or capacity-based). We also explored the sensitivity of the results to the time of the viewing session (first half, second half, or at the end considering the session in its entirety). We identified a set of four hypotheses based on subjective quantitative, continuous subjective quantitative, and qualitative measures. We also evaluated a novel continuous subjective measurement technique.

The findings surrounding hypotheses 1 and 2 suggest that a buffer-based strategy might provide a better experience under higher network impairment conditions. For the two network scenarios considered, the buffer-based strategy is effective in avoiding stalls but does so at the cost of reduced video quality (based on the lower $playerRate$ results). However, the buffer-based strategy does yield a lower number of quality switches, see Table IV, as a result of infrequent bitrate adaptations which have been shown to negatively affect user perceived quality [Cranley et al. 2006; Ni et al. 2011]. Participants

12http://cybertiger.clemson.edu/vss/userStudy.html
in Strategy 1 do notice the drop in video quality causing a decrease in perceived QoE, but the perceived levels of video quality, viewer frustration, and opinions of video clarity and distortion are significantly worse due to artifacts such as stalls in Strategy 2, as compared to Strategy 1. The capacity-based strategy tries to provide the highest video quality possible but produces many more artifacts during playback. The results suggest that player video quality has more of an impact on perceived quality when stalls are infrequent.

The quantitative subjective results did not support the third hypothesis. However as these measures were recorded using a post-survey after the session, it is possible that one or more particular artifacts observed by the participants may have swayed their opinion. The same might not have been true for the continuous perceived quality rating. Another reason for the above observation could be that the participants started to get accustomed to the artifacts and did not react to them as readily later during the video playback.

The objective of our study was to provide design guidance for a HAS adaptation algorithm and to explore a continuous measurement technique. Gathering from our results, we put forward a comparison of 2 different streaming approaches along with data that is used to analyze how viewers respond to different artifacts. We also believe that our novel continuous measurement technique can be used to accurately record QoE scores for future experiments. Based on these findings, we put forward some guidelines for developers and service providers alike; 1) using a streaming technique that is a compromise of the two techniques presented in this paper is most likely the best approach; 2) segmenting video content into smaller high quality chunks that can provide a faster rate of transfer may reduce the likelihood of buffer stalls; 3) a client side player that switches between the two strategies based on the network bandwidth available may also improve the QoE for viewers.

A limitation of our work is that we did not test these strategies on smaller mobile platforms and the results may not transfer to other media viewing platforms. Another limitation is that we did not investigate QoE for video content in other genres which may not include as many visual effects as the current video and may perhaps produce significantly different reactions. In ongoing work, we continue to explore and understand the human reaction to specific artifacts (or sequences of artifacts) that we have collected. We also plan to consider the impact of both sound quality and network impairment on the viewer’s comprehension of the plot and storyline of the short movie under each of the strategies. The objective would be to empirically evaluate if the presence and types of artifacts have cognitive effects on viewers. Finally, we plan on extending the scope and scale of the user study by developing and conducting an online study designed to obtain results from potentially a very large number of participants.

9. ACKNOWLEDGEMENT

This research was supported in part by CableLabs, Comcast, and Intel. We thank our colleagues from Clemson University who provided insight and expertise that greatly assisted the research.

REFERENCES


ACM Transactions on Applied Perception, Vol. 0, No. 0, Article 0, Publication date: August 2018.


S. Cook, B. Mathieu, P. Truong, and I. Hamchaoui. 2017. QUIC: Better for what and for whom?. In 2017 IEEE International Conference on Communications (ICC), 1–6. DOI:http://dx.doi.org/10.1109/ICC.2017.7997281


ACM Transactions on Applied Perception, Vol. 0, No. 0, Article 0, Publication date: August 2018.
Comparative Evaluation of User Perceived Quality Assessment of Design Strategies for HTTP-Based Adaptive Streaming


