Enabling Perception-Driven Optimization in Networking

Yihua Cheng, Xu Zhang, Junchen Jiang (University of Chicago)
{yihua98,zhangxu,junchenj}@uchicago.edu

ABSTRACT

Service providers struggle to catch up with the rapid growth in bandwidth and latency demand of Internet videos and other applications. An essential contributor to this resource contention is the assumption that users are equally sensitive to service quality everywhere, so any low-quality incidents must be avoided. However, this assumption is not true. For example, our work and other parallel efforts have shown that more video users can be served with better quality of experience (QoE) if we embrace the fact that the QoE's sensitivity to video quality varies greatly with the video content. To unleash such benefits, the application systems must be driven by not only system measurement data but also user feedback data that capture users' perceptions of service quality. In this short paper, I will highlight some of our recent efforts toward the efficient collection of user feedback and enabling perception-driven optimization for Internet applications.

1. INTRODUCTION

The landscape of online applications has seen a sea change over the past few years, with multiple trends driving up the bandwidth demands for online videos. The rise of ultra high-definition (4K/8K/VR) videos dramatically increases the per-video bandwidth demand and is projected to be 22% of global video traffic in 2022 from 3% in 2017. Video traffic to mobile devices has also more than tripled in the last 3 years. The rising bandwidth demands widen the gap between user expectation and user-perceived *quality of experience* (*QoE*) measured in mean opinion score or user engagement.

Achieving better bandwidth-QoE tradeoffs relies on accurate QoE models. Widely used in modern video delivery systems, a QoE model takes a streamed video (such as buffering stalls and visual quality index, etc) as input and returns a predicted QoE as output. Most adaptive-bitrate (ABR) and CDN/ISP resource allocation algorithms (*e.g.,* [9, 6]) use QoE models to predict when increasing video quality has more QoE improvement. Thus, any errors of a QoE model can mislead these optimization techniques to pick suboptimal decisions and miss opportunities to improve QoE or save bandwidth.

Indeed, recent efforts have shown that the sensitivity of QoE to these optimizations differs significantly across videos, web pages, and even across different segments of the same video (*e.g.,* [11, 2]). Therefore, having more accurate QoE measurements allows content providers to strategically allocate more compute/bandwidth resources or enhance quality at points of higher QoE sensitivity (see §2). With these trends, QoE measurements are increasingly needed.

While there have been many efforts to make QoE measurements faithfully reflect true user experience, relatively less attention has been given to *building a system that obtains QoE measurements fast.* Two relevant efforts exist—one automates QoE measurements by using crowdsourcing and the other uses collected QoE measurements to dynamically prune videos that no longer need QoE ratings. Unfortunately, it is challenging to combine the two ideas, because with the existing crowdsourcing interface, one must specify which videos to be rated by how many users *before* each crowdsource task begins, making it hard to dynamically prune redundant videos without launching multiple crowdsourcing campaigns. In short, prior work suffers from two limitations: *(i)* The speed to obtain QoE measurements is still quite slow due to the traditional crowdsourcing interface; and *(ii)* QoE measurements can be obtained for only on-demand content, not live content.

This short paper introduces two projects[1] that aim at addressing these limitations. First, to speed up QoE crowdsourcing, we have *developed and open-sourced[2] VidPlat, the first re-usable tool for fast and automated QoE measurements.* VidPlat allows dynamic pruning of QoE video samples in one single crowdsource task. To realize it, VidPlat creates a new shim layer between the researchers and the crowdsourcing platform, allowing researchers to define a logic that iteratively creates new videos that need more ratings based on the latest QoE measurements. Compared to existing QoE measurement methods, VidPlat (1) keeps all QoE measurements in one crowdsourcing task, thus minimizing the overhead to initialize tasks and re-calibrate/train raters, (2) dynamically decides when enough ratings are gathered for each video, thus reducing the total number of QoE ratings, and (3) is an open-source platform that future researchers can re-use and customize.

Second, to enable the QoE measurements on live videos, we present SensitiFlow, an alternative architecture that *online* profiles and adapts to quality sensitivity by continuously gathering and analyzing QoE-related feedback from real video sessions watching the same video. Concretely, SensitiFlow orchestrates the adaptive-bitrate (ABR) logic of video sessions. SensitiFlow runs an online *feedback loop* with two components. At a high level, SensitiFlow maintains the quality-sensitivity profile of each video segment by continuously collecting QoE-related feedback from ongoing video sessions, which can be average user ratings or whether a higher percentage of users quit/skip watching a video segment under low quality than under high quality. Based on the up-to-date quality-sensitivity profiles, SensitiFlow makes ABR decisions to improve QoE for concurrent and future video sessions watching the same video.

2. WHY IMPROVING EFFICIENCY OF QOE MEASUREMENTS

Since it is hard to directly ask users to rate their subjective

[1] `https://people.cs.uchicago.edu/~junchenj/perception_driven_optimization`
[2] `https://github.com/orgs/QoEStudies/repositories`

Copyright is held by author/owner(s).

experience in real-time, researchers and content providers run offline user studies to assess QoE under various objective quality metrics. Participants are asked to watch an application *demo*. In video QoE, a demo can be a video streamed with a one-second buffering stall deliberately added at a certain point. In web QoE, a demo can be a web page loaded with a certain page load time (*e.g.,* a certain above-the-fold time). Then the participants rate the subjective QoE score in the range from 1 to 5. Finally, we can calculate the mean QoE scores of each demo video and model the relationship between QoE and quality metrics.

Potential of QoE-driven optimization: Traditionally, QoE models are expected to capture the *general* relationship between QoE and a few quality metrics. As a result, once enough QoE measurements are collected to model QoE on several representative videos or web pages, the QoE models will be re-used on other videos or pages. However, many recent efforts have shown that more granular, *context-specific* QoE models, which quantify the QoE-quality relationship of *individual* video (or even video segments) [11]. The shift from one-size-fits-all QoE models to context-specific QoE models quickly increases the frequency and amount of QoE measurements. For instance, Netflix produces on average more than 580 minutes worth of new video content every day. If it builds a separate QoE model for each minute of video, it will ask 10 raters to watch and rate 100 hours of videos every day. Similarly, web QoE research also shows a similar increase in the demand for QoE measurements [2, 3]. These context-specific QoE models can substantially improve QoE without using more bandwidth or compute resources. For example, in video streaming, applying per-video QoE models to adaptive bitrate (ABR) algorithms in video players can improve 15.4% QoE without using more network bandwidth [11]; in web services, we can have 40% QoE improvement by allocating computing resources across different web requests by their QoE models [3, 12]. More QoE measurements are also needed when a new optimization (*e.g.,* a new video bitrate ladder or chunk segmentation [7]) is proposed whose impact on QoE may not be captured by existing QoE models.

Related work and limitations: As the need for QoE measurements rises, so does the need to reduce the *latency* (to recruit workers and collect QoE ratings) and the *cost* of QoE measurements (total compensation given to the workers who provide the QoE ratings).

Two efforts exist to reduce the delay and cost of QoE measurements. First, several efforts (*e.g.,* [5, 10]) have shown the potential of automating QoE measurements using **crowdsourcing platforms** like Amazon Mechanical Turk (MTurk) and Prolific. These works have been focused on retaining reliable crowd workers, calibrating QoE ratings, mitigating hidden confounders (*e.g.,* order of assignment completion or different user devices), and reducing cost via dynamic pricing. Second, depending on the QoE ratings already collected, many demos would be **redundant** and can be **pruned** to let participants rate fewer demos [8]. For instance, to investigate how video bitrate affects the QoE of a particular video, if human raters are unable to perceive the QoE difference between bitrates of 1 Mbps and 10 Mbps, then no ratings will be needed for the bitrates between 1 Mbps and 10 Mbps on this video.

A natural question then is *will QoE measurement be automated and made much faster as promised by these approaches?* Unfortunately, the answer is no, because to dynamically prune demos, researchers have to *sequentially launch a series of crowdsourcing tasks* and use the QoE measurements from one task to decide which demo can be pruned in the next task, causing significant delays.

3. VidPlat: A PLATFORM FOR FASTER QOE CROWDSOURCING

To fully realize the speed benefit of crowdsourced QoE measurements, we have developed *VidPlat, the first re-usable open-source tool that enables dynamic demo pruning to speed up crowdsourced QoE measurements.* VidPlat serves as a shim layer between the researchers and the crowdsourcing platforms. VidPlat launches one task, but unlike the traditional crowdsourcing interface that requires researchers to pre-determine the demos and the number of QoE ratings per demo upfront, VidPlat offers a more flexible interface to researchers. Researchers are allowed to define a logic and a few initial demos, and upon receiving a QoE rating, VidPlat invokes this logic to determine the subsequent demos based on the logic's output. Then instead of immediately showing the next demo to raters, the next demo will be first put in a queue, and VidPlat will decide which demo in the queue should be given to the next rater. Using this "indirection" between which demos need more QoE ratings and which demo to be rated next by a worker, VidPlat retains the flexibility to randomize the order demos seen by a rater and avoid asking a worker to rate too many (similar) demos.

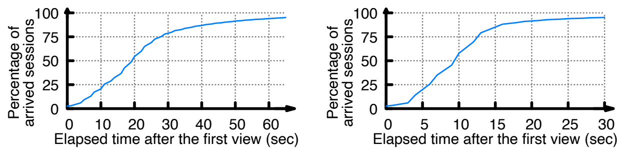
In short, with VidPlat, researchers do not need to determine all the demos or the required number of QoE ratings before the user study task begins; instead, VidPlat lowers the development burden while still collecting crowdsourced QoE measurements with minimum redundancy. As a result, it greatly reduces the number of demos and QoE ratings collected, thereby saving both time and cost.

Use cases: VidPlat has already been used in three IRB-approved QoE-related projects: *(i)* investigating the relationship between webpage load time and QoE [12]; *(ii)* exploring the correlation between video quality and QoE in on-demand video streaming [11]; and *(iii)* comparing the QoE impact of video bitrate and motion-to-photon (MTP) latency in online video gaming [4]. VidPlat's dynamic assignment determination significantly improved the efficiency of our user studies. For instance, compared to Sensei [11], a prior tool employing a traditional interface, VidPlat reduced both costs and latency by more than 50% in these use cases, while obtaining QoE models that realize the same QoE improvement as Sensei. These empirical results demonstrate the tangible benefits of our novel approach.

4. SensitiFlow: ENABLING QOE-DRIVEN OPTIMIZATION IN LIVE VIDEOS

Though VidPlat speeds up QoE measurements, the content of the video must be known *beforehand*. How to estimate the variability of quality sensitivity in *live* videos? One may use heuristics, such as VMAF or the popularity of a video segment to infer quality sensitivity, but how sensitive users are to quality under different content is often more complex than what can be captured by these heuristics [11].

So can we use real users to profile quality sensitivity, especially in live videos? Fortunately, this is feasible since most views of a live video segment occur dur-



(a) A (live-linear) TV show (b) A live-event sports video

Figure 1: *Example arrival patterns of views of the same live video segment: 20% of sessions watch the same content at least 3 seconds earlier than 60% of sessions.*

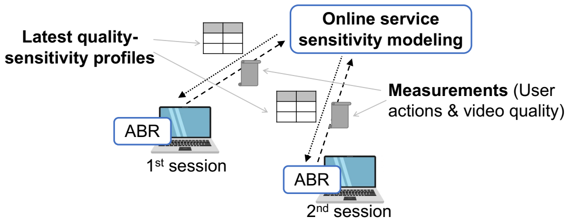


Figure 2: *Each session in SensitiFlow uses the latest quality-sensitivity profile to make ABR decisions and updates the global coordinator with the latest user actions.*

ing a non-trivial time span of 30-40 seconds after it first being viewed. Figure 1 shows the relative wall-clock time of sessions watching a chunk of a live-linear video (24/7 live programs [1] such as talk shows, TV plays) and a live-event video (*e.g.,* live sports broadcasting). Our conversation with domain experts has confirmed that such time discrepancies among live video viewers are commonly accepted in the industry. Over the last decades, the modest time difference (10-30 seconds) among viewers has become an accepted feature (rather than a bug) of live internet videos and efforts to realize full synchronicity in large-scale live events have been lukewarm due to the implementation complexity.

Inspired by this observation, SensitiFlow's *global coordinator* (depicted in Figure 2) constantly collects online measurements of per-segment quality and QoE-related feedback (*e.g.,* exit or skip) from video sessions to maintain an up-to-date view of the *quality-sensitivity profile* of each video. A quality-sensitivity profile maps each segment and quality level to the estimation and variance of quality sensitivity, in engagement drops and retention drops. It can answer the "what-if" question: *what would the expected drop in engagement/retention for a given quality at each segment?* When a session's ABR logic decides the bitrate of the next video segment, it will query the quality-sensitivity profiles and make ABR decisions using logic such as the one described in [11].

To test the gains of SensitiFlow on user engagement (QoE), we consider a simple logic that works in two phases. In the *profiling phase*, the first N sessions use a *default ABR logic*, and their per-segment user engagement and quality metrics are collected and used to estimate the quality sensitivity of the segment. After N sessions, it enters the *optimization phase*, in which each session runs a variant of the quality-sensitivity-aware ABR algorithm proposed in [11]. The algorithm takes as input the player's current state (history throughput, buffer length, etc) and the quality sensitivity of the next three segments and returns as output ABR decision for the next chunk. We evaluated QoE in user engagement (view time) using real traces of 7.6 million video sessions from a content provider. Our preliminary results show that SensitiFlow can realize up-to 80% of the improvement obtained by a hypothetical "oracle" system that knows quality sensitivity in advance.

5. VISION: USER-CENTRIC NETWORKING

SensitiFlow and VidPlat show the early promise of a more *user-centric* approach, where *measurements on user experience and actions* are first-class citizens of system monitoring and optimization. Just like systems metrics indicate current system states, user actions and engagement reveal individual user's experiences, as they watch a video, browse a web page, or use a mobile app. While we study only video systems in this paper, we think that the general approach may be applicable to other network applications such as gaming and mobile web. For instance, users' tolerance to web page loading time is better modeled by directly observing users' natural actions (*e.g.,* [12, 10]). This user-centric approach calls for novel system designs to realize the *tight* control loop between (near-)real-time user experience measurement and system adaptation. SensitiFlow and VidPlat take a step in this direction, and working toward such a perception-driven system is an active direction of future research.

6. REFERENCES

[1] Linear Live Streaming 101. `https://antmedia.io/linear-live-streaming-101/`.

[2] M. Butkiewicz, D. Wang, Z. Wu, H. V. Madhyastha, and V. Sekar. Klotski: Reprioritizing web content to improve user experience on mobile devices. In *NSDI*, 2015.

[3] P. Casas, S. Wassermann, N. Wehner, M. Seufert, and T. Hossfeld. Not all web pages are born the same content tailored learning for web qoe inference. In *2022 IEEE International Symposium on Measurements & Networking (M&N)*, pages 1–6. IEEE, 2022.

[4] Y. Cheng, A. Arapin, Z. Zhang, Q. Zhang, H. Li, N. Feamster, and J. Jiang. Grace++: Loss-resilient real-time video communication under high network latency. *arXiv preprint arXiv:2305.12333*, 2023.

[5] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel. Best practices and recommendations for crowdsourced qoe-lessons learned from the qualinet task force" crowdsourcing". 2014.

[6] J. Jiang, S. Sun, V. Sekar, and H. Zhang. Pytheas: Enabling data-driven quality of experience optimization using group-based exploration-exploitation. In *USENIX NSDI*, 2017.

[7] M. Licciardello, L. Humbel, F. Rohr, M. Grüner, and A. Singla. [solution] prepare your video for streaming with segue. *Journal of Systems Research*, 2(1), 2022.

[8] X. Liu, W. Song, Q. He, M. Di Mauro, and A. Liotta. Speeding up subjective video quality assessment via hybrid active learning. *IEEE Transactions on Broadcasting*, 69(1):165–178, 2022.

[9] H. Mao, R. Netravali, and M. Alizadeh. Neural adaptive video streaming with pensieve. In *ACM SIGCOMM*, 2017.

[10] M. Varvello, J. Blackburn, D. Naylor, and K. Papagiannaki. Eyeorg: A platform for crowdsourcing web quality of experience measurements. In *CoNEXT*, 2016.

[11] X. Zhang, Y. Ou, S. Sen, and J. Jiang. Sensei: Aligning video streaming quality with dynamic user sensitivity. In *USENIX NSDI*, 2021.

[12] X. Zhang, S. Sen, D. Kurniawan, H. Gunawi, and J. Jiang. E2e: embracing user heterogeneity to improve quality of experience on the web. In *ACM SIGCOMM*. 2019.