published in *Journal of Digital Video*, vol. 4, no. 1, pp. 52-57, Dec. 2019.

Encoding Intelligence for Optimal Viewer Experience in Live Video Distribution

Letter to the Editor prepared for SCTE•ISBE
by

Zhou Wang, Professor, University of Waterloo
Chief Science Officer, SSIMWAVE Inc.
200 University Ave W, Dept. of ECE, University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada
zhou.wang@uwaterloo.ca
519-888-4567 ex. 35301

Abdul Rehman, CEO, SSIMWAVE Inc.
375 Hagey Boulevard, Suite 310
Waterloo, Ontario, N2L 6R5, Canada
abdul.rehman@ssimwave.com
519-489-2688

Kai Zeng, Lead Researcher, SSIMWAVE Inc.
375 Hagey Boulevard, Suite 310
Waterloo, Ontario, N2L 6R5, Canada
kai.zeng@ssimwave.com
519-489-2688

1. Introduction

Real-world live video distribution systems are often faced with the great challenge of processing videos of extremely diverse content type and complexity. The challenge becomes even greater given the critical real-time requirement and the large volume of 24/7 video streams that need to be processed. Using a fixed encoding setup to drive the live video encoders for bandwidth reduction, as is the case in most real-world live distribution systems, causes serious problems, resulting in encoded/transcoded videos that often suffer from severe and unpredictable quality variations across time, video assets, and content types.

In the case of live video distribution, decisions need to be made instantaneously to make the best options for encoder configurations easily adopted in the video encoding/transcoding pipeline.

To empower the encoder with intelligence requires two key components:

1. A quality-of-experience (QoE) metric that not only accurately predicts end viewers experience when consuming videos streamed to their viewing devices, but is also real-time and light-weight, producing consistent QoE predictions across content type, content complexity, codec type, bit rate, video resolution, frame rate and dynamic range; and

2. An intelligent optimization engine that drives the encoders to produce the best and controllable QoE scores in diverse environment and meanwhile maximizing bandwidth reduction.

Working solutions that best address these critical issues are highly desirable for live video distributions.

2. User Experience Metrics for Encoding Performance

An objective user QoE metric aims to automatically predict end viewer's visual experience when watching the encoded video fully decoded and rendered on their viewing devices. Objective QoE assessment is a difficult task because it requires deep understanding about how the sophisticated encoding process creates compression artifacts for diverse types of video content and how such artifacts impact the quality assessment behavior of the human visual system (HVS). Traditionally a direct numerical measure, namely the peak signal-to-noise-ratio (PSNR), has been commonly used for encoder evaluation and comparison, but PSNR has been shown to have low correlation with perceived video quality [1]. There has been a great deal of effort in the past two decades developing advanced objective metrics that better predict subjective video quality. Representative metrics include the structural similarity index (SSIM) [1], [2], the multi-scale SSIM (MS-SSIM) [3], the information content-weighted SSIM (IW-SSIM) [4], the video quality model (VQM) [5] and the video multi-method assessment fusion (VMAF) [6]. These metrics demonstrate significantly improved video quality predictions under certain controlled test conditions. Nevertheless, they are still highly limited in terms of their functionality, interpretability, application scope, and computational cost. Such limitations often make it extremely difficult, if not completely impossible, to use these objective metrics in various real-world video encoding/transcoding scenarios, especially in time-critical applications such as live video distributions. In recent years, novel objective QoE metrics designed to overcome these problems are emerging. These metrics target two types of crucial properties, which will be elaborated here.

The first type of properties focus on the accuracy, speed, cost and interpretability of the QoE metric. There is no doubt that the QoE metric should produce video quality scores that accurately predict viewer experiences. The standard way to test the accuracy of an objective metric is to compute the linear correlation coefficient, rank-order correlation coefficient, and mean prediction error, between the objective scores and

mean subjective opinions using large-scale subject-rated video databases. The metric also needs to have low computational and implementation cost, readily deployed in large-scale video distribution systems. This will also allow for high-speed computation for continuous 24/7 real-time assessment of high-resolution, high frame rate and high dynamic range videos with moderate hardware configurations. The metric must also be easily interpretable, producing quality scores that linearly relate to what an average viewer would say about the quality of a video. For example, if the quality score range may be between 0 and 100, divided into five evenly spaced segments corresponding to five perceptual QoE levels of bad (0-20), poor (21-40), fair (41-60), good (61-80), and excellent (81-100) quality, respectively. Such a metric creates an easy-to-grasp common language, allowing smooth communication in large organizations, where engineers and operators can identify and fix quality problems on the fly, researchers and developers can optimize individual components and the overall video delivery systems, and executives can make critical business decisions.

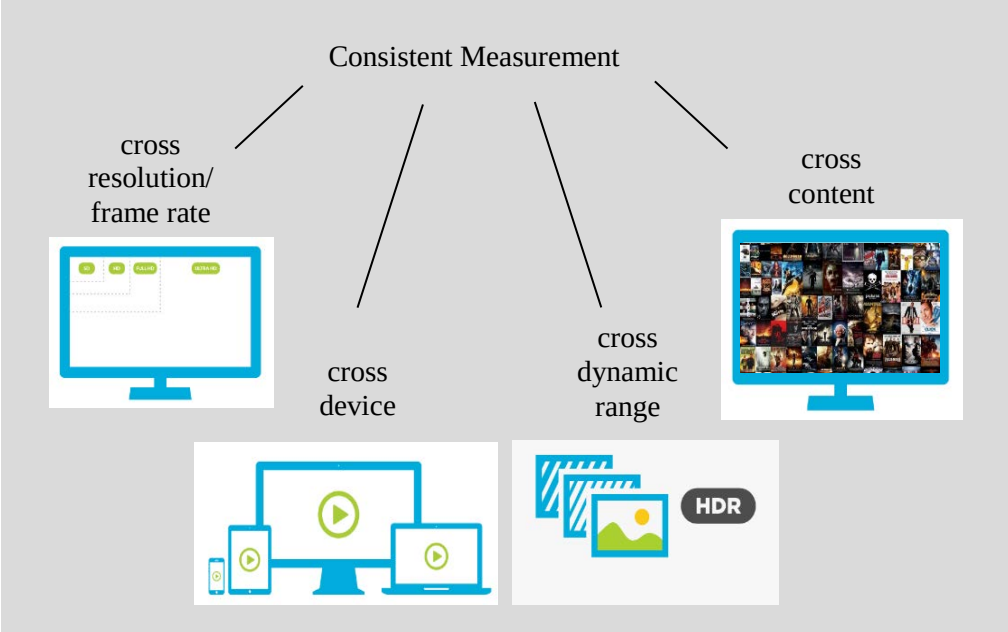

Figure 1 – Critical requirements lacking in traditional QoE metrics

The second type of critical properties relate to the usability and consistency of the QoE metric in real-world application scenarios. It is important to note that well-known video quality metrics (PSNR, SSIM, MS-SSIM, IW-SSIM, VQM, VMAF) require pixel-to-pixel correspondence between the reference and test videos. As a consequence, when videos at the input and output of the video encoder/transcoder are of different spatial resolutions, frame rates, and dynamic ranges, these metrics often do not apply. This greatly impedes the practical usage of these metrics because in modern video distribution, it is very common that the source input videos are transcoded into multiple versions of not only different bit rates, but also different spatial resolutions, frame rates and dynamic ranges. In addition, the playbacks of the same video stream on different viewing devices could create significantly different viewer experiences, but these metrics often generate one quality score only (or a few scores corresponding to a few different devices), and thus fail to capture the device variations of visual QoE assessment. Another common but important issue with these quality metrics is that they often create inconsistent scores across content of different types and complexity

levels. As a result, scores generated by these metrics cannot be compared across content, meaning that two videos of similar perceptual QoE may be given drastically different scores, largely constraining the practical use such QoE metrics in large-scale distribution systems that make instantaneous resource allocation decisions across hundreds or thousands of video services and live video channels. Therefore, as shown in Figure 1, in real-world video distribution systems, it is essential to use a QoE metric that simultaneously produces consistent quality measurements across spatial resolutions, frame rates, dynamic ranges, viewing devices, and video content.

Recently, great effort has been made to develop novel QoE metrics for the above-mentioned properties. So far, the full SSIMPLUS Viewer Score metric is offering all these critical properties [7],[8], and the open source VMAF project has also been making progress towards the direction [9].

3. Encoding Intelligence Driven by User Experience Metrics

A good QoE metric that satisfies the critical properties is a fundamental ingredient to enable encoding intelligence. On top of that, an encoding decision-making engine driven by content and encoding performance analysis may be used to control the live encoding/transcoding process. This may be done in different ways, and two types of encoding intelligence frameworks are described below.

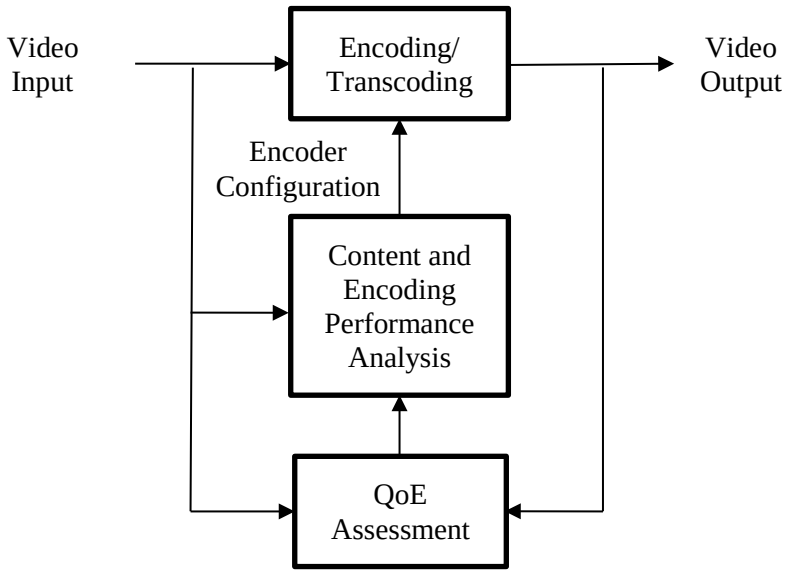

Figure 2 – Type I Encoding Intelligence

The first type of encoding intelligence works for the application scenarios where the encoder or transcoder configurations can be controlled on-the-fly. These configurations may include the spatial and temporal resolutions, the bit rate, the quantization parameter (QP), the group-of-picture (GoP) structure, the encoding pre-set, and other parameters that may influence the encoding process. When the source input video is received, it first goes through content analysis that may include spatial, temporal and color complexity measurement, content type analysis, dynamic range and color statistics, and other statistics of the content. Meanwhile, the QoE metric, which compares the current input and output video streams before and after the encoder/transcoder, is computed and then fed into the analysis module. Based on both content and

encoding performance analysis, decisions on encoder/transcoder configurations are made and used to control the encoder/transcoder instantaneously. The intelligence decisions should be geared towards the best balancing point between sustained quality delivery and cost-effective bandwidth usage. This process is illustrated in Figure 2.

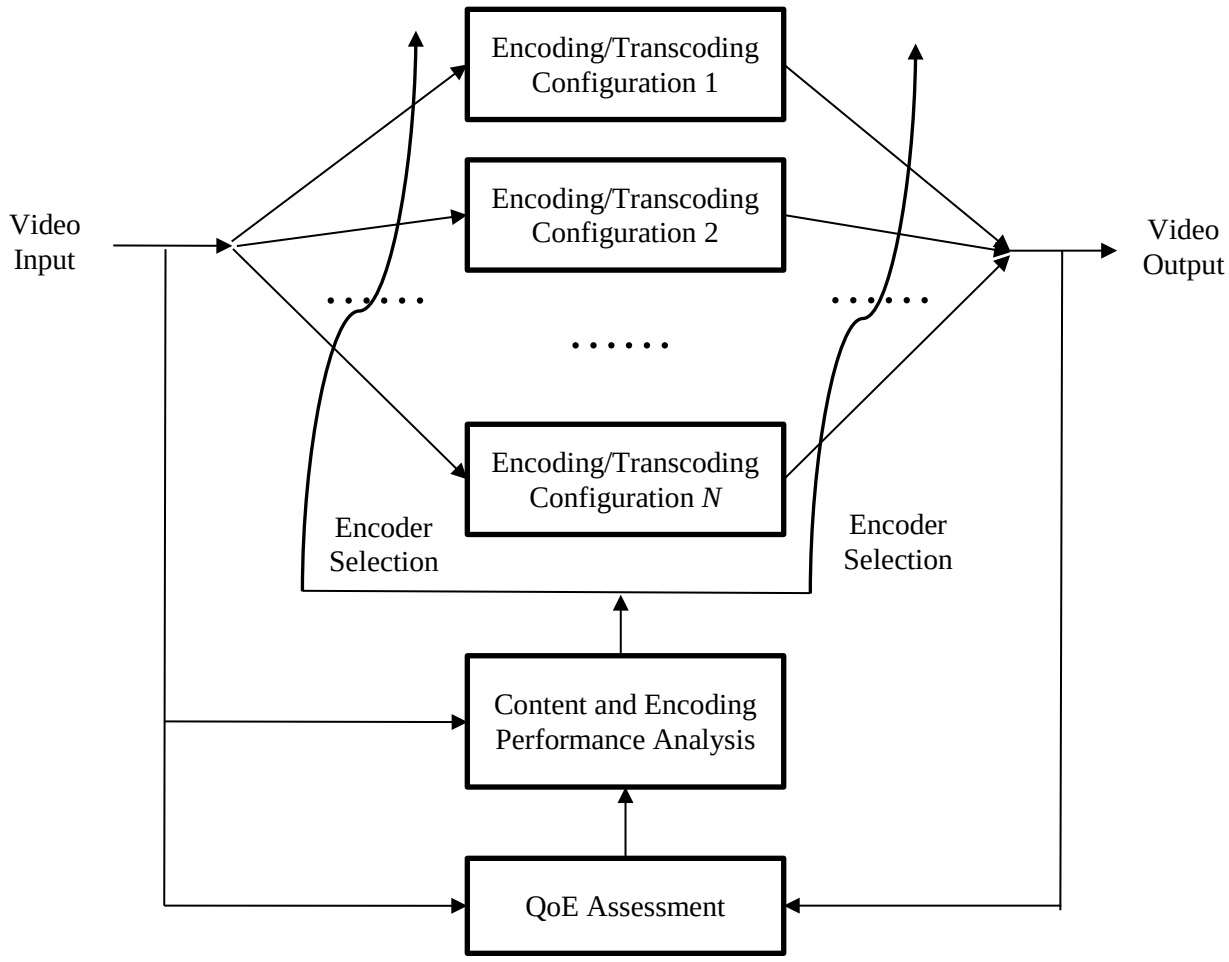

Figure 3 – Type II Encoding Intelligence

The second type of encoding intelligence adapts to the scenarios where on-the-fly encoder parameter adjustment is difficult, but multiple encoder/transcoder configurations are setup previously. As a result, the intelligence is on the selection of encoders from multiple options, as shown in Figure 3. The pre-determined encoder/transcoder configurations may be designed to target at videos of different content types and spatial/temporal/color complexity levels. They could also represent different types of encoding technologies or encoder solutions. Similar to the Type I intelligence case, source content analysis is performed and the QoE metric between the current input and output video streams before and after the encoder/transcoder is computed instantaneously. Both types of information is employed by the content and encoding performance analysis module to create an intelligence decision that choses one out of the multiple encoder/transcoder configuration options for the next step or encoding event.

In both encoding intelligence frameworks, each encoder/transcoder block may be designed to generate one output video stream or a ladder of outputs (which includes multiple encoded videos of different resolutions, frame rates, and bit rates), depending on the deployment points in the video delivery chain and also on the specific use cases. In addition, the analysis and decision-making processes may be based on either short-term instantaneous inputs, or on long-term statistics.

4. Conclusions

Compared with video-on-demand (VoD) and many other use cases, encoding intelligence for live video distribution is more challenging because all the critical decisions need to be made instantaneously, any suboptimal decisions need to be identified and corrected on-the-fly, and the solutions need to work robustly and continuously 24/7 in large-scale systems. The tolerance of errors is often low, and any wrong decision may lead to severe and unpredictable quality issues, immediately affecting a large number of end viewers' visual experiences [10]. The two most crucial components for encoding intelligence is the QoE metric and the encoding intelligence engine. We discussed the challenges and state-of-the-art solutions for both components. We have also discussed two types of general frameworks on how QoE-driven encoding intelligence may be deployed in real-world application scenarios.

5. Bibliography and References

[1] *Image quality assessment: from error visibility to structural similarity*, Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, IEEE Transactions on Image Processing, Apr. 2004.

[2] *Video quality assessment based on structural distortion measurement*, Z. Wang, L. Lu, and A. C. Bovik, *Signal Processing: Image Communication*, Feb. 2004.

[3] *Multi-scale structural similarity for image quality assessment*, Z. Wang, E. P. Simoncelli and A. C. Bovik, IEEE Asilomar Conference on Signals, Systems and Computers, Nov. 2003.

[4] *Information content weighting for perceptual image quality assessment*, Z. Wang and Q. Li, IEEE Transactions on Image Processing, May 2011.

[5] *A new standardized method for objectively measuring video quality*, M. H. Pinson, IEEE Transactions on Broadcasting, Sept. 2004.

[6] *Toward a practical perceptual video quality metric*, Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy and M. Manohara, Netflix TechBlog, Jun 2016.

[7] *Display device-adapted video quality-of-experience assessment*, A. Rehman, K. Zeng and Z. Wang, IS&T/SPIE Electronic Imaging: Human Vision & Electronic Imaging, Feb. 2015.

[8] *SSIMPLUS: The most accurate video quality measure*, https://www.ssimwave.com/from-the-experts/ssimplus-the-most-accurate-video-quality-measure/

[9] *VMAF: the journey continues*. Z. Li. C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy and J. De Coc, Netflix TechBlog, 2019.

[10] *Begin with the end in mind: a unified end-to-end quality-of-experience monitoring, optimization and management framework*, Z. Wang and A. Rehman, SMPTE Motion Imaging Journal, 2019.