Perceptually-aware Live VBR Encoding Scheme for Adaptive Video Streaming

Vignesh V Menon
Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität, Klagenfurt, Austria

13 Dec 2022

Outline

Introduction

HTTP Adaptive Streaming (HAS)[1]

Why Adaptive Streaming?

- Adapt for a wide range of devices.
- Adapt for a broad set of Internet speeds.

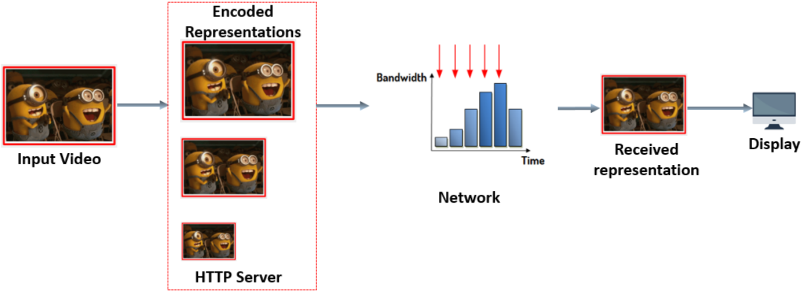
What HAS does?

- Each source video is split into segments.
- Encoded at multiple bitrates, resolutions, and codecs.
- Delivered to the client based on the device capability, network speed *etc.*

[1] A. Bentaleb et al. "A Survey on Bitrate Adaptation Schemes for Streaming Media Over HTTP". In: *IEEE Communications Surveys Tutorials* 21.1 (2019), pp. 562–585. DOI: 10.1109/COMST.2018.2862938.

Introduction

HTTP Adaptive Streaming (HAS)



- *HTTP Adaptive Streaming* (HAS) has become the *de-facto* standard in delivering video content for various clients regarding internet speeds and device types.
- Traditionally, a fixed bitrate ladder, *e.g.*, *HTTP Live Streaming* (HLS) bitrate ladder[2], is used in live streaming.

[2] https://developer.apple.com/documentation/http_live_streaming/ hls_authoring_specification_for_apple_devices, last access: Dec 12, 2022.

Introduction

Motivation for Per-title bitrate ladder in Adaptive Streaming

Each resolution performs better than others in a specific region for a given bitrate range. These regions depend on the *video content complexity*.

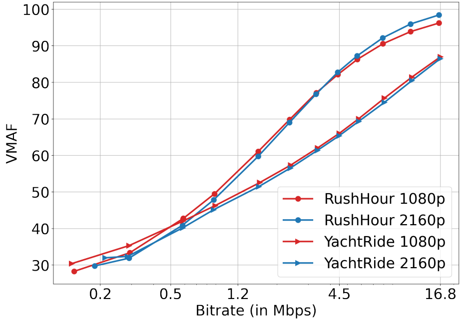


Figure: Rate-Distortion (RD) curves of the Constant Bitrate (CBR) encoding of *RushHour_s000* and *YachtRide_s000* video sequences (segments) of VCD dataset[3] encoded at 1080p and 2160p resolutions using x265 HEVC encoder at *ultrafast* preset. Here, VMAF is used as the quality metric.

[3] Hadi Amirpour et al. "VCD: Video Complexity Dataset". In: *Proceedings of the 13th ACM Multimedia Systems Conference*. MMSys '22. Athlone, Ireland: Association for Computing Machinery, 2022, 234–239. ISBN: 9781450392839. DOI: 10.1145/3524273.3532892. URL: https://doi.org/10.1145/3524273.3532892.

Introduction

Per-title Encoding

- Though per-title encoding schemes[4,5,6] enhance the quality of video delivery, determining the *convex-hull* is computationally costly, making it suitable for only VoD streaming applications.
- The plethora of live streaming applications call for low latency approaches to optimize video coding.
- According to the Bitmovin Video Developer Report 2021[7], *live (low) latency* is the biggest challenge in video technology today.

[4] Jan De Cock et al. "Complexity-based consistent-quality encoding in the cloud". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 1484–1488. DOI: 10.1109/ICIP.2016.7532605.

[5] Madhukar Bhat, Jean-Marc Thiesse, and Patrick Le Callet. "A Case Study of Machine Learning Classifiers for Real-Time Adaptive Resolution Prediction in Video Coding". In: *2020 IEEE International Conference on Multimedia and Expo (ICME)*. 2020, pp. 1–6. DOI: 10.1109/ICME46284.2020.9102934.

[6] Daniel Silhavy et al. "Machine Learning for Per-Title Encoding". In: *SMPTE Motion Imaging Journal* 131.3 (2022), pp. 42–50. DOI: 10.5594/JMI.2022.3154836.

[7] https://go.bitmovin.com/video-developer-report, last access: Dec 13, 2022.

Introduction

Motivation for perceptually-aware bitrate ladder

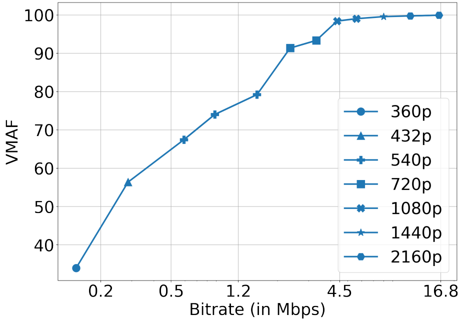


Figure: RD curve of the HLS CBR encoding of *Characters_s000* video sequence (segment) of VCD dataset using x265 HEVC encoder at *ultrafast* preset. The points with a bitrate greater than 3.6 Mbps are in the perceptually lossless region. Hence, there is significant storage wastage while storing these representations.

Selecting similar-quality representations for the bitrate ladder does not result in improved QoE, but it will lead to increased storage and bandwidth costs!

Introduction

Motivation for two-pass encoding (CBR versus VBR)

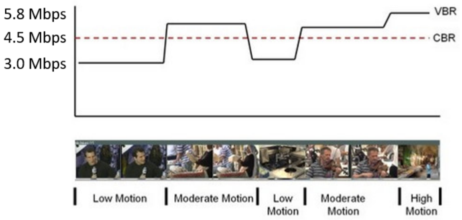


Figure: Constant Bitrate (CBR) versus Variable Bitrate (VBR) encoding.

- In live streaming, Constant Bitrate (CBR) rate-control mode is used to encode video sequences at a fixed bitrate ladder. The consistency of CBR makes it more reliable for time-sensitive data transport.
- In this method, there is no concern about the bitrate exceeding internet speeds. However, this method may result in low compression efficiency.

Introduction

Constrained Variable Bitrate (cVBR) encoding

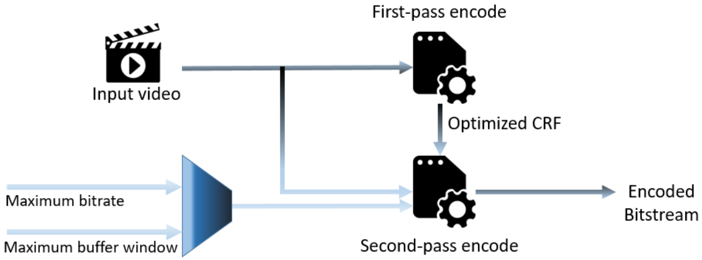


Figure: cVBR encoding.

- A "rate factor" first-pass to identify the optimized CRF to achieve the target bitrate.
- In the second pass, the segment is encoded with the selected optimized CRF with the maximum bitrate and maximum buffer window constraints.
- The desired target bitrate is achieved with **maximum compression efficiency and minimum quality fluctuation**.

Research Goal

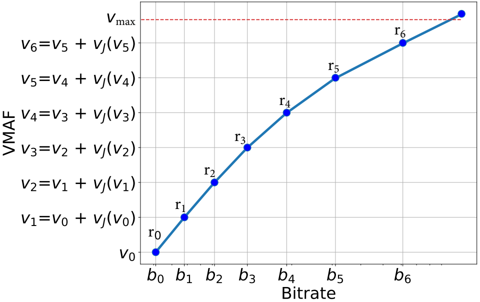


Figure: The ideal perceptually-aware bitrate ladder envisioned in this work. Here, $v_J(v_0) = v_J(v_1) = v_J(v_{M-1}) = \Delta\text{VMAF}$

Joint optimization:

- Perceptual difference of pre-defined ΔVMAF between representations.
- Minimize bitrate difference between representations.
- Maximize compression efficiency of representations.

Workflow of Live-VBR

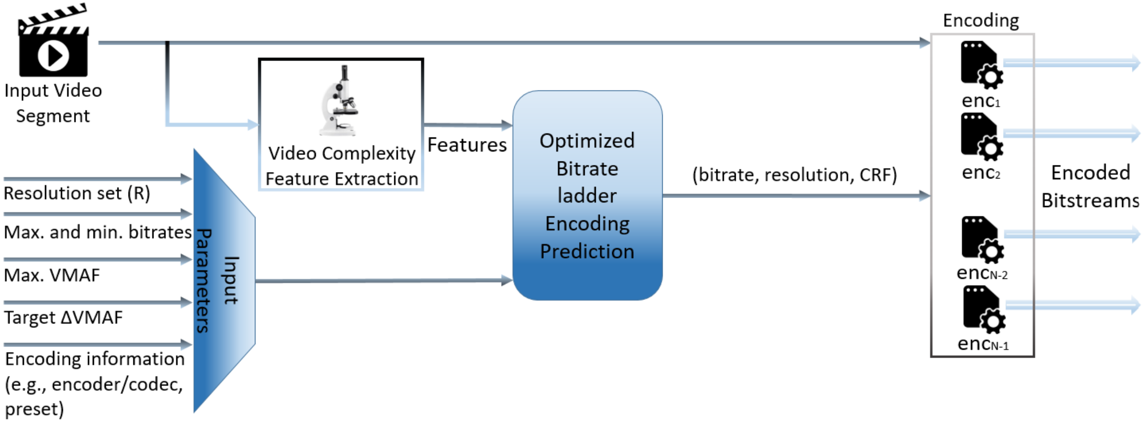


Figure: Live-VBR system envisioned in this work.

Video Complexity Feature Extraction

Compute texture energy per block

A DCT-based energy function is used to determine the block-wise feature of each frame defined as:

$$H_k = \sum_{i=0}^{w-1} \sum_{j=0}^{w-1} e^{|(\frac{ij}{wh})^2 - 1|} |DCT(i,j)| \quad (1)$$

where wxw is the size of the block, and $DCT(i,j)$ is the $(i,j)^{th}$ DCT component when $i + j > 0$, and 0 otherwise.

The energy values of blocks in a frame are averaged to determine the energy per frame.[8,9]

$$E_s = \sum_{k=0}^{K-1} \frac{H_{s,k}}{K \cdot w^2} \quad (2)$$

[8] Michael King, Zinovi Tauber, and Ze-Nian Li. "A New Energy Function for Segmentation and Compression". In: *2007 IEEE International Conference on Multimedia and Expo*. 2007, pp. 1647–1650. DOI: 10.1109/ICME.2007.4284983.

[9] Vignesh V Menon et al. "Efficient Content-Adaptive Feature-Based Shot Detection for HTTP Adaptive Streaming". In: *2021 IEEE International Conference on Image Processing (ICIP)*. 2021, pp. 2174–2178. DOI: 10.1109/ICIP42928.2021.9506092.

Video Complexity Feature Extraction

h_s: SAD of the block level energy values of frame s to that of the previous frame $s-1$.

$$h_s = \sum_{k=0}^{K-1} \frac{|\, H_{s,k}, H_{s-1,k} \,|}{K \cdot w^2} \tag{3}$$

where K denotes the number of blocks in frame s.
The luminescence of non-overlapping blocks k of s^{th} frame is defined as:

$$L_{s,k} = \sqrt{DCT(0,0)} \tag{4}$$

The block-wise luminescence is averaged per frame denoted as L_s as shown below.[10]

$$L_s = \sum_{k=0}^{K-1} \frac{L_{s,k}}{K \cdot w^2} \tag{5}$$

[10] Vignesh V Menon et al. "VCA: Video Complexity Analyzer". In: *Proceedings of the 13th ACM Multimedia Systems Conference*. MMSys '22. Athlone, Ireland: Association for Computing Machinery, 2022, 259–264. ISBN: 9781450392839. DOI: 10.1145/3524273.3532896. URL: https://doi.org/10.1145/3524273.3532896.

Live-VBR

First point of the bitrate ladder[11]

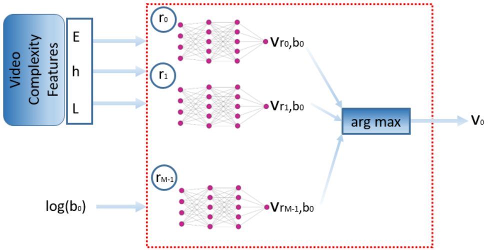


Figure: Estimation of the first point of the bitrate ladder. v_0 is the maximum value among the v_{r,b_0} values output from the predicted models trained for resolutions r_0, r_1, .., r_{M-1}. The resolution corresponding to the VMAF v_0 is chosen as r_0.

$b_0 = b_{min}$
Determine v_{r,b_0} $\forall r \in R$
$v_0 = max(v_{r,b_0})$
$r_0 = \arg\max_{r \in R}(v_{r,b_0})$
(r_0, b_0) is the first point of the bitrate ladder

Note

This part of the algorithm needs VMAF prediction for all considered resolutions.

[11] V. V. Menon et al. "OPTE: Online Per-Title Encoding for Live Video Streaming". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 1865–1869. DOI: 10.1109/ICASSP43922.2022.9746745.

Live-VBR

Remaining points of the bitrate ladder

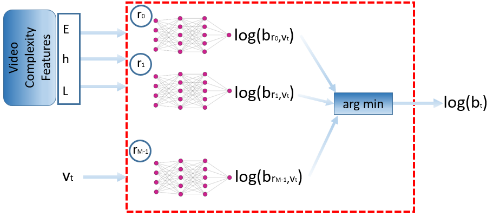


Figure: Estimation of the $(t+1)^{th}$ point of the bitrate ladder. b_t is the minimum value among the b_{r,v_t} values output from the predicted models trained for resolutions r_0, r_1, .., r_{M-1}. The resolution corresponding to the bitrate b_t is chosen as r_t.

$t = 1$
for $t \geq 1$ **do**
  $v_t = v_{t-1} + \Delta\text{VMAF}$
  Determine b_{r,v_t} $\forall r \in R$
  $b_t = min(b_{r,v_t})$
  $r_t = \arg\min_{r \in R}(b_{r,v_t})$
  if $b_t > b_{max}$ *or* $v_t > v_{max}$ **then**
    End of the algorithm
  else
    (r_t, b_t) is the $(t+1)^{th}$ point of the bitrate ladder.
    $t = t + 1$

Note

This part of the algorithm needs bitrate prediction for all considered resolutions.

Live-VBR

cVBR encoding of the bitrate ladder[12]

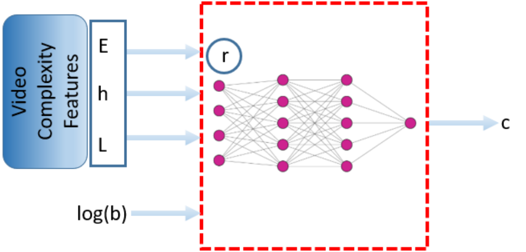


Figure: Estimation of the optimized CRF to achieve the target bitrate b using a prediction model trained for resolution r.

- Optimized CRF is determined for the selected (r, b) pairs.
- cVBR encoding for the (r, b, CRF) pairs is performed.

[12] Vignesh V Menon et al. "ETPS: Efficient Two-Pass Encoding Scheme for Adaptive Live Streaming". In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 1516–1520. DOI: 10.1109/ICIP46576.2022.9897768.

Results

Prediction accuracy of the models

Table: R^2 score and MAE of the prediction models for various resolutions.

	R^2 score							MAE						
r	360p	432p	540p	720p	1080p	1440p	2160p	360p	432p	540p	720p	1080p	1440p	2160p
VMAF	0.821	0.852	0.882	0.906	0.910	0.906	0.930	4.860	4.899	4.832	4.393	3.838	3.490	2.941
log(b)	0.859	0.864	0.888	0.915	0.932	0.937	0.943	0.765	0.751	0.737	0.709	0.711	0.706	0.681
CRF	0.969	0.969	0.970	0.969	0.968	0.967	0.965	1.924	1.920	1.914	1.942	1.940	1.972	1.990

Note

Just three values (E, h, L) are used as the measure of video complexity. If we increase the information measure, *e.g.*, block-wise features), the accuracy can be improved further.

Results

RD plots of Live-VBR using x265

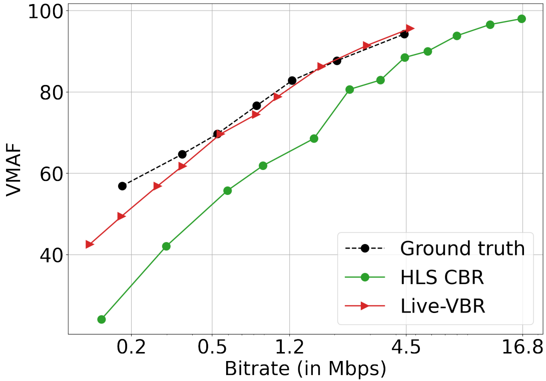


Figure: *Bunny_s000*$(E = 22.40,\ h = 4.70)$

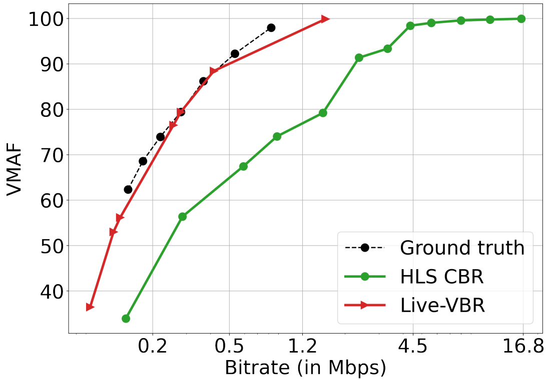


Figure: *Characters_s000*$(E = 45.42,\ h = 36.88)$

Results

RD plots of Live-VBR using x265

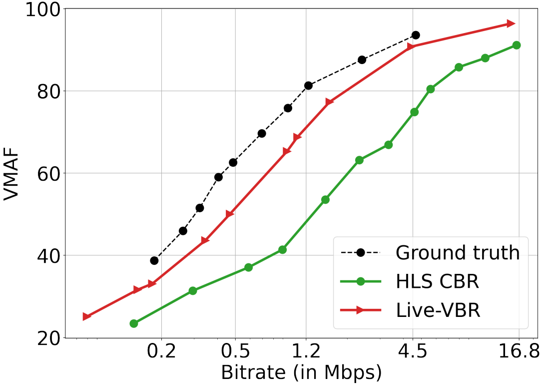


Figure: *Eldorado_s005* $(E = 100.37,\ h = 9.23)$

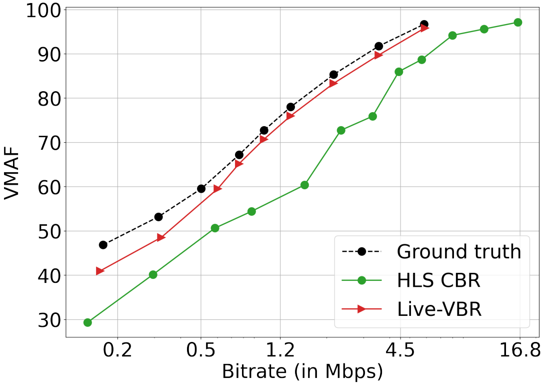


Figure: *Wood_s000* $(E = 124.72,\ h = 47.03)$

Results

Summary

Table: Average results of the encoding schemes compared to the HLS CBR encoding using x265 HEVC encoder.

Method	BDR_P	BDR_V	BD-PSNR	BD-VMAF	ΔS	ΔE
Ground truth (ΔVMAF=2)	-23.09%	-43.23%	1.34 dB	10.61	-25.99%	89.54%
Ground truth (ΔVMAF=4)	-28.15%	-42.75%	1.70 dB	10.08	-59.07%	-0.54%
Ground truth (ΔVMAF=6)	-25.36%	-40.73%	1.67 dB	9.19	-70.50%	-31.24%
Live-VBR (ΔVMAF=2)	-14.25%	-29.14%	1.36 dB	7.82	23.57%	90.19%
Live-VBR (ΔVMAF=4)	-18.41%	-32.48%	1.41 dB	8.31	-56.38%	0.34%
Live-VBR (ΔVMAF=6)	-18.80%	-32.59%	1.34 dB	8.34	-68.96%	-28.25%

Relative storage difference

$$\Delta S = \frac{\sum b_{opt}}{\sum b_{ref}} - 1$$

Relative energy utilization difference

$$\Delta E = \frac{\sum E(b_{opt})}{\sum E(b_{ref})} - 1$$

Summary and Future Directions

- Presented an application of video complexity analysis, where VMAF, target bitrate, and CRF are predicted using video complexity features.

In the future, we shall include the following:

- Optimized encoding framerate
- Optimized encoding preset and number of CPU threads

Q & A

Thank you for your attention!

Vignesh V Menon (vignesh.menon@aau.at)