Between Two and Six?
Towards Correct Estimation of JND Step Sizes for VMAF-based Bitrate Laddering

ATHENA

Hadi Amirpour [1] Raimund Schatz [2] Christian Timmerer [1]

[1]Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität, Klagenfurt, Austria
[2]AIT Austrian Institute of Technology, Austria

UNIVERSITÄT KLAGENFURT

Overview

- Video streaming platforms such as Netflix, YouTube and Amazon Prime Video have become integral part of our daily lives, in particular after COVID-19 crisis.
- *HTTP Adaptive Streaming* (HAS) is the prevailing technique for both live and *Video on Demand* (VoD) streaming applications.
- In HAS, each video content is encoded at multiple bitrate-resolution pairs (or quality-resolution pairs), referred as to *representations*, to construct a bitrate ladder.
- Providing representations with different quality levels in a bitrate ladder enables the dynamic matching of video quality to end-user's available bandwidth and device type.
- Bitrate ladders are typically optimized per content using per-title encoding approaches.

Per-title encoding

- Each video content is encoded at multiple bitrates and resolutions and a convex hull is formed based on the quality of encodings.
- Since **VMAF** yields the highest performance in predicting the quality of video stream, it is widely used to evaluate quality of encodings.
- Bitrate-resolution pairs are selected from the convex hull to construct an optimized bitrate ladder.

Question?

Which encodings to select from the convex hull to construct a bitrate ladder?

Just Noticeable Difference (JND)

- The HVS is capable of differentiating only a few discrete-scale distortion levels in a wide range of bitrates in a compressed video.
- The minimum visual difference that can be perceived by HVS, *i.e.*, the difference between two adjacent perceptual distortion levels, refers as to one ***Just Noticeable Difference (JND)***.
- The first JND point denotes the transitional point from perceptually lossless to perceptually lossy coding.
- Selecting encodings with noticeable quality differences in between prevents the construction of an inefficient bitrate ladder that suffers from too similar quality representations.

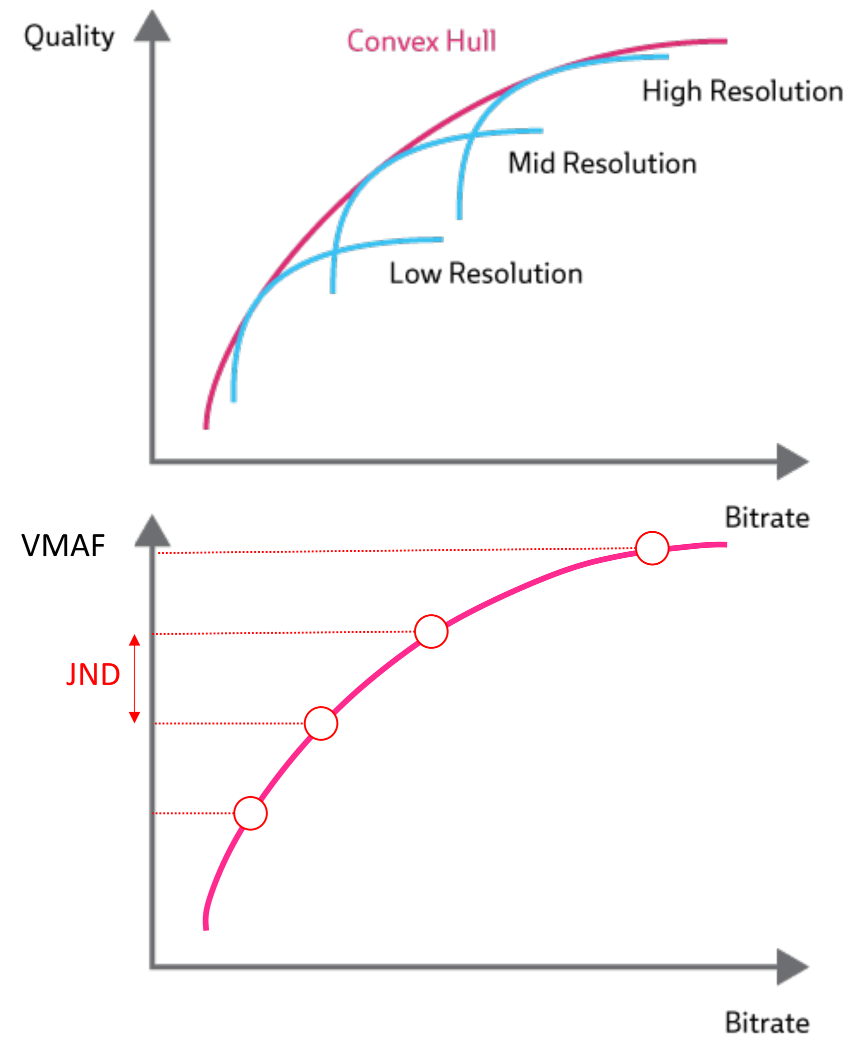

Figure 1. JND in per-title encoding.

Question?

For a given video content clip, how to determine JND-optimal step sizes for efficient bitrate laddering in the VMAF domain?

- We used a large-scale JND-based video quality dataset, named VideoSet, containing 220 source video sequences with 5s duration.
- Videos are encoded with the constant quantization parameter (CQP) rate control mode of H.264/AVC in QP range of [0,51].
- In VideoSet, one JND step refers to the distortion level where SUR is equal to 75%, *i.e.*, 75% of user can distinguish the distortion between two representations. The subjective tests were conducted to find the QP boundaries of the 1st, 2nd, and 3rd JNDs.
- Two sources provide concrete recommendations for sizing ΔVMAF:
  — Jan Ozer [2] recommends ΔVMAF $=6$.
  — Kah et al. [1] recommend ΔVMAF $=2$.

The huge variance (sd=3.334) of ΔVMAF as depicted in Fig. 2 shows that ***there is no simple rule of thumb for JND-optimal ΔVMAF***, since optimal step size varies considerably from clip to clip.

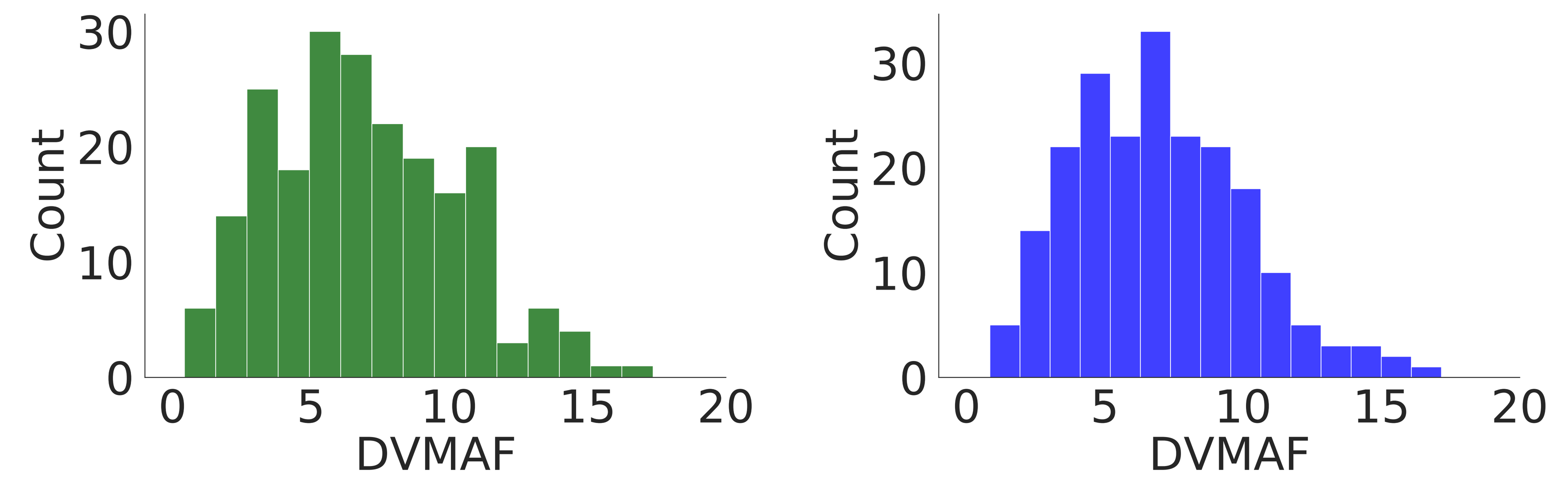

Figure 2. Distributions of ΔVMAF values in the dataset. Left: ΔVMAF between all adjacent JND point pairs. Right: ΔVMAF between JND points 2 & 3 only as example.

— Six frame-wise features including: (1) Spatial Information (SI), (2) Temporal Information (TI), (3) Spatial Energy (E), (4) Temporal Energy (h), (5) Brightness (L), and (6) Colourfulness (c) are extracted from the original video and in addition to (7) Frame rate (fr) are used to represent the characteristics of videos.
— We found that a GLM with feature selection based on lasso regularization ($\alpha = 0.01$) provided the best fit with the data.
— We found that it is sufficient to calculate them for the uncompressed source clips only.

Table 1. Evaluation results for the different ΔVMAF step-size estimation models.

Model	RMSE	MAE	R^2
ΔVMAF = 2	5.962	5.008	-2.232
ΔVMAF = 6	3.451	2.743	-0.083
ΔVMAF = 6.93	3.316	2.726	0.000
GLM	2.649	2.110	0.362

Table 2. GLM coefficients for the features used.

E (mean)	3.544	TI (mean)	-1.052
h (median)	3.469	c (mean)	-0.644
SI (mean)	-3.159	fr	-0.501
L (median)	2.364		

Acknowledgment

The financial support of the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology, and Development, and the Christian Doppler Research Association is gratefully acknowledged. Christian Doppler Laboratory ATHENA: `https://athena.itec.aau.at/`.

References

[1] Andreas Kah, Christopher Friedrich, Thomas Rusert, Christoph Burgmair, Wolfgang Ruppel, and Matthias Narroschke. Fundamental relationships between subjective quality, user acceptance, and the VMAF metric for a quality-based bit-rate ladder design for over-the-top video streaming services. In Andrew G. Tescher and Touradj Ebrahimi, editors, *Applications of Digital Image Processing XLIV*, page 38, San Diego, United States, August 2021. SPIE.

[2] Jan Ozer. Finding the Just Noticeable Difference with Netflix VMAF. https://www.linkedin.com/pulse/finding-just-noticeable-difference-netflix-vmaf-jan-ozer/, 2017.