HAL open science

A Real-Time Video Quality Metric for HTTP Adaptive Streaming

Hadi Amirpour, Jingwen Zhu, Patrick Le Callet, Christian Timmerer

To cite this version:

Hadi Amirpour, Jingwen Zhu, Patrick Le Callet, Christian Timmerer. A Real-Time Video Quality Metric for HTTP Adaptive Streaming. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr 2024, Seoul, France. pp.3810-3814, 10.1109/ICASSP48485.2024.10446839 . hal-04615893

HAL Id: hal-04615893

https://hal.science/hal-04615893v1

Submitted on 18 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Autorisation HAL

HAL Authorization

A REAL-TIME VIDEO QUALITY METRIC FOR HTTP ADAPTIVE STREAMING

Hadi Amirpour[1], Jingwen Zhu[2], Patrick Le Callet[2], Christian Timmerer[2]

[1]Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität, Klagenfurt, Austria
[2]Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, IUF, Nantes, France

ABSTRACT

In HTTP Adaptive Streaming (HAS), a video is encoded at multiple bitrate-resolution pairs, referred to as representations, which enables users to choose the most suitable representation based on their network connection. To optimize the set of bitrate-resolution pairs and improve the Quality of Experience (QoE) for users, it is of utmost importance to measure the quality of the representations. VMAF is a highly reliable metric used in HAS to assess the quality of representations. However, in practice, using it for optimization can be a very time-consuming process, and it is infeasible for live streaming applications. To tackle its high complexity, our paper introduces a new method called *VQM4HAS*, which extracts *low-complexity* features, including (i) video complexity features, (ii) bitstream features logged during the encoding process, and (iii) basic video quality metrics. These extracted features are then fed into a regression model to predict VMAF. Our experimental results demonstrate that *VQM4HAS* achieves a high Pearson Correlation Coefficient (PCC) with VMAF, ranging from 0.95 to 0.96 depending on the resolution. However, it exhibits significantly lower complexity, making it suitable for live streaming scenarios.

Index Terms— Video quality, HAS, VMAF, QoE, bitstream.

1. INTRODUCTION

Video streaming is experiencing a steady increase in usage, thanks to the widespread availability of the high-speed internet and the proliferation of mobile devices [1]. It has become an integral part of our daily lives, with applications ranging from entertainment to education, business, and beyond. For example, streaming platforms such as Netflix, Hulu, and Amazon Prime Video give us the ability to watch our favorite content anytime, anywhere. Online learning platforms, like Coursera and Udemy, provide students with the flexibility to learn at their own pace and from anywhere in the world, using video streaming technology. Additionally, video streaming technology enables real-time communication between people regardless of their location, resulting in significant time and cost savings by eliminating the need for travel.

Video streaming services typically rely on *HTTP Adaptive Streaming* (HAS) [2] as their primary technology for delivering content to viewers. HAS divides each video content into smaller segments, enabling clients to receive and play the video immediately without waiting for the entire video to download. Furthermore, each video segment is provided at multiple bitrates, known as representations, allowing clients to adjust dynamically based on the available network bandwidth and the device's capabilities, ensuring a smooth viewing experience [3, 4].

Measuring the perceived video quality [5] is an important aspect of evaluating QoE in video streaming services [6, 7]. Video quality is typically affected by compression artifacts that occur during video compression, which is necessary to reduce the bitrate of the video for efficient transmission over networks. The video compression process uses codecs to encode the video data and remove redundant information, resulting in the loss of some details and visual quality.

The most accurate method of assessing video quality is subjective testing, which is very time-consuming and expensive [8, 9]. As a cost-effective alternative to subjective testing, objective metrics can be used to evaluate video quality. Among objective quality metrics, VMAF [10] has gained significant attention in recent years due to its high correlation with subjective testing results [11, 12, 13]. While VMAF is highly accurate in predicting human video quality perception, it has been criticized for its slow processing speed [14]. This can in-

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/.

crease computation costs for content providers and make it impractical for use in real-time applications. This is particularly important for streaming applications, where the quality metric needs to be evaluated multiple times for the same video, across different representations.

In this paper, we propose a novel video quality metric named *VQM4HAS* to reduce the computational demand for the calculation of video quality in HAS. The proposed *VQM4HAS* metric significantly reduces the computational cost of VMAF while maintaining a high correlation with that. This is achieved by using a combination of low-complexity features, such as (i) video complexity features, (ii) bitstream features *logged during encoding*, and (iii) basic video quality metrics logged during encoding, for making predictions.

2. A VIDEO QUALITY METRIC FOR HAS

In this section, we present our proposed video quality metric, *VQM4HAS*, which utilizes low-complexity features extracted from the original video and video encodings to incorporate them for prediction. These features are classified into three categories: (i) video complexity features (cf. 2.1), (ii) bitstream features (cf. 2.2), and (iii) basic quality metrics (cf. 2.3) which are used to construct *VQM4HAS*, as shown in Fig. 1.

2.1. Video Complexity Features

The perceived quality of video frames at a fixed bitrate is affected by various factors, including the spatial and temporal complexities of the frames. These features have a significant impact on video compressibility [15] and are commonly utilized in video quality metrics [16]. The spatial complexity of frames refers to the amount of detail and variation in the video frames, such as textures, edges, and colors. Temporal complexity, on the other hand, refers to the amount of motion and changes between frames, such as camera movements, object movements, and scene changes. Videos with high spatial and temporal complexity typically require higher bitrates to maintain the same perceived video quality. Therefore, we compute video complexity features for both the original and encoded videos, denoted as C_{org} and C_{rec}, respectively. In this paper, we use a DCT-based energy function (E_{DCT}) [17, 18] to quantify the spatial complexity of video frames. The energy function serves as a tool for mapping the texture of a block of pixels from a multi-dimensional frequency space to a one-dimensional energy space. It assigns higher costs to higher DCT frequencies. The DC value is treated separately. The function is defined as follows:

$$E_{DCT} = \sum_{i=1}^{w} \sum_{j=1}^{h} e^{\left[\left(\frac{ij}{wh}\right)^2 - 1\right]} |DCT(i-1, j-1)| \quad (1)$$

where E_{DCT} represents the energy of the block, with w and h denoting its width and height, respectively. The $(i, j)^{th}$ DCT component is given by $DCT(i, j)$ when $i + j$ is greater than 2, and is 0 otherwise. The energy function per frame E is calculated by averaging over entire blocks [18]. Blocks of highly textured pixels produce high energy values, as they contain many high-frequency DCT values. Conversely, blocks of homogeneously colored pixels, which have few high-frequency DCT values, produce low energy values.

For the temporal complexity, we use the temporal function of the DCT-based energy function (h) [18], which is expressed as follows:

$$h = \frac{1}{C} \sum_{c=1}^{C-1} \frac{1}{w^2} SAD(E_{DCT}(t, c), E_{DCT}(t-1, c)) \quad (2)$$

here, c denotes the block index, C represents the total number of blocks, w signifies the width of the blocks, and t denotes the frame index. In summary,

$$C_{rec} = \{E_{rec}, h_{rec}\}, C_{org} = \{E_{org}, h_{org}\}. \quad (3)$$

2.2. Bitstream Features

The encoding statistics and information are valuable data because they can represent the compressibility of the video. We selected some features from the bitstream log during encoding as follows:

- ***QP*** (Quantization Parameter) is a parameter used in video coding to control the quantization process, which in turn affects the bitrate and video quality. In the constant bitrate rate control method, QP is dynamically determined for each block so that the bitrate reaches the target level. Lower QP values result in better video quality but higher bitrate, while higher QP values result in lower video quality but lower bitrate. The average QP per frame is recorded as a feature during video encoding.
- ***Bits***, which denotes the number of bits required to encode each frame, is another feature that can be logged during encoding. This feature can provide information on the complexity of each frame, which can be useful for predicting video quality.
- ***Distortion*** denotes the average distortion video frames after compression and is typically reported as two separate values: one for the luma (brightness) component and one for the chroma (color) component of the frames.

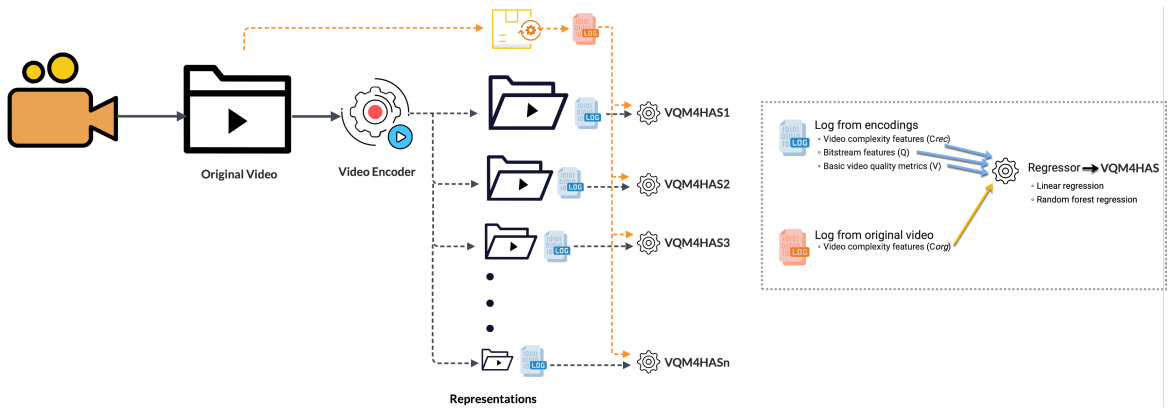


Fig. 1: Overview of *VQM4HAS* architecture.

- ***Psy Energy*** refers to the amount of energy in a video frame that is not perceived by the human visual system. It is calculated as the sum of the absolute difference (SAD) between the source and the reconstructed energy of a frame.
- ***Residual Energy*** is a measure of the energy of the difference between the original image and the reconstructed image. It is calculated as the sum of squared error (SSE) between the original image and the reconstructed image before quantization.
- ***Luma/Chroma Values*** represent the brightness (luma) and color (chroma) information of a video frame. For our evaluation, we focused solely on their average values and disregarded their minimum and maximum values.
- ***Total CTU Time*** is the average time for compressing and filtering Coding Tree Units (CTUs) [19] of each frame

In summary,

$$B = \{qp, bit, luma_dist, chroma_dist, psy_energy \\ res_energy, luma_avg, cr_avg, cb_avg, CTU_time\} \quad (4)$$

2.3. Basic Video Quality Metrics

Video codecs often provide two quality metrics, namely *PSNR* and *SSIM*. These can be enabled by appending options like `--psnr` and `--ssim` to the FFmpeg command line. These metrics, denoted as Q in this paper, can be calculated during encoding without incurring a significant additional time cost compared to performing a separate calculation, as the encoder computes them on-the-fly. This eliminates the need to invoke external software and read frames again, thereby improving efficiency.

3. EXPERIMENTAL RESULTS

In this section, we present the experimental results by evaluating *VQM4HAS* on the Inter4K dataset [20]. It contains 1000 ultra-high (4K) resolution video clips with a frame rate of 60 frames per second (fps) sourced from YouTube. The videos cover a variety of content types. We use the open-source software x265 v3.4 to encode videos, following the recommendations of Apple [21] by producing 12 representations of the HLS bitrate ladder. The spatial (E_{org}) and temporal (h_{org}) complexity features are extracted from the original videos using the open-source VCA software [18]. The features of the encoded representation are extracted by logging them into a csv file using the `--csv-log-level` option in x265. The x265 encoder was modified to log the spatial (E_{rec}) and temporal (h_{rec}) complexity features during encoding. To compute VMAF scores, the encoded representations are first decoded and then upscaled to a 4K resolution. The VMAF score is then calculated between the original video and the upscaled representation using the VMAF model vmaf_4k_v0.6.1[1].

3.1. Performance Analysis

To predict the per-segment VMAF scores for each representation in the bitrate ladder, we perform temporal pooling and calculate the average of the per-frame features for each segment. We then use both linear regression and random forest models to predict the scores. Table 2 summarized the PCC scores obtained when using *VQM4HAS* to predict per-segment VMAF. It is observed that the PCC of VMAF is slightly lower only for the lowest bitrate (0.87), whereas with increasing bitrate, the PCC can reach as high as 0.96.

[1] https://github.com/Netflix/vmaf/blob/master/resource/doc/models.md

Table 1: HLS bitrate ladder.

	Represnetation ID	1	2	3	4	5	6	7	8	9	10	11	12
Bitrate ladder	**Bitrate (kbps)**	145	300	600	900	1600	2400	3400	4500	5800	8100	11600	16800
	Resolution	360p	432p	540p	540p	540p	720p	720p	1080p	1080p	1440p	2160p	2160p

Table 2: PCC for *VQM4HAS* when predicting per-segment VMAF scores.

	Represnetation ID	1	2	3	4	5	6	7	8	9	10	11	12
Model	**Linear**	0.83	0.86	0.90	0.91	0.93	0.95	0.95	0.95	0.95	0.94	0.92	0.90
	Random Forest	0.87	0.90	0.94	0.94	0.95	0.95	0.95	0.96	0.95	0.95	0.95	0.94

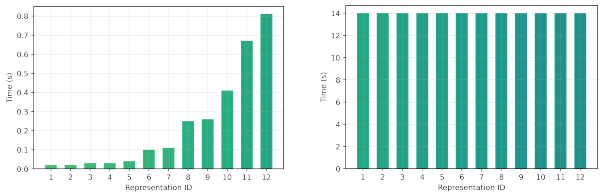
In live video streaming, a fixed bitrate ladder is typically used [22], which makes it feasible to use a regression model for each representation. However, if a new bitrate-resolution pair is added to the ladder, a new regression model will be required. Considering that infinite bitrates are possible for encoding, we evaluate *VQM4HAS* per resolution, as only a limited number of standard resolutions are used for encoding. To achieve this, we train a random forest model on the training set of representations that share the same resolution. For instance, we train a random forest model on the training videos with bitrates of 600 kbps, 900 kbps, and 1600 kbps, all having a resolution of 540p. This model is used to predict the VMAF scores of the test set for the same representations. The PCC results for resolutions that have more than one bitrate in the HLS ladder (Table 1) are summarized in Table 3.

Table 3: PCC for *VQM4HAS* when predicting per-resolution VMAF scores.

Resolution	540p	720p	1080p	2160p
VMAF prediction	0.95	0.96	0.96	0.96

3.2. Time-complexity Analysis

In this section, we compare the time complexity of VMAF with the proposed *VQM4HAS*. To this end, we calculate all the metrics on an Amazon EC2 c5.4xlarge instance assuming that the videos are being watched on a 4K display. Therefore, for the VMAF calculation, the representations are upscaled to 4K, which is also the resolution of the original video. This means that the time complexity of the VMAF calculation is independent of the resolution of the representation, while *VQM4HAS* is dependent on the bitrate and resolution of the representation. The results indicate that *VQM4HAS* exhibits significantly lower time complexity in comparison to VMAF, making it suitable even for live streaming applications. The maximum delay added by the computation of *VQM4HAS* is related to the representation #12 in the bitrate ladder, which has a delay of 2.5 ms. This delay is much lower than that of the real-time computation (33.3 ms or 30 fps).



(a) (b)

Fig. 2: (a) The time complexity of different methods when computing on 12 representations of a 5-second video encoded with the HLS ladder parameters. The complexity breakdown across representations for (a) *VQM4HAS*, (b) VMAF

4. CONCLUSION

This paper introduced a new video quality metric called *VQM4HAS*, which is capable of accurately predicting VMAF with high correlation while maintaining significantly lower computational complexity. In order to accurately model perceived video quality, low complexity features, including (i) video complexity features, (ii) bitstream features, and (iii) basic video quality metrics, are extracted and fed into a regression model. All of these features are logged during the encoding of representations, incurring minimal costs, and only the video complexity features are extracted from the original video. Linear and random forest regression models were evaluated for the study. Per-resolution evaluations demonstrated that *VQM4HAS* is capable of predicting VMAF with a Pearson correlation coefficient (PCC) ranging from 0.95 to 0.96. The prediction of VMAF is largely influenced by a few key features depending on the resolution/representation. This makes it highly suitable for use in live streaming applications.

5. REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Forecast and Trends, 2018–2023," *White Paper*, Mar. 2020.

[2] Iraj Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," *IEEE Multimedia*, vol. 18, no. 4, pp. 62–67, Apr. 2011.

[3] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[4] Babak Taraghi, Minh Nguyen, Hadi Amirpour, and Christian Timmerer, "Intense: In-Depth Studies on Stall Events and Quality Switches and Their Impact on the Quality of Experience in HTTP Adaptive Streaming," *IEEE Access*, vol. 9, pp. 118087–118098, 2021.

[5] Guangtao Zhai and Xiongkuo Min, "Perceptual Image Quality Assessment: A Survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 211301, Nov. 2020.

[6] Hadi Amirpour, Christian Timmerer, and Mohammad Ghanbari, "PSTR: Per-Title Encoding Using Spatio-Temporal Resolutions," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, July 2021, pp. 1–6, ISSN: 1945-788X.

[7] Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer, "DeepStream: Video Streaming Enhancements using Compressed Deep Neural Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[8] Margaret H Pinson and Stephen Wolf, "Comparing Subjective video Quality Testing Methodologies," in *Visual Communications and Image Processing 2003*. SPIE, 2003, vol. 5150, pp. 573–582.

[9] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan C Bovik, and Lawrence K Cormack, "A Subjective Study to Evaluate Video Quality Assessment Algorithms," in *Human Vision and Electronic Imaging XV*. SPIE, 2010, vol. 7527, pp. 128–137.

[10] Netflix Technology Blog, "VMAF: The Journey Continues," Oct. 2018.

[11] Reza Rassool, "VMAF reproducibility: Validating a perceptual practical video quality metric," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Cagliari, Italy, June 2017, pp. 1–2, IEEE.

[12] Nabajeet Barman, Steven Schmidt, Saman Zadtootaghaj, Maria G. Martini, and Sebastian Möller, "An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming," in *Proceedings of the 23rd Packet Video Workshop*, Amsterdam Netherlands, June 2018, pp. 7–12, ACM.

[13] C. Lee, S. Woo, S. Baek, J. Han, J. Chae, and J. Rim, "Comparison of Objective Quality Models for Adaptive Bit-Streaming Services," in *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Larnaca, Aug. 2017, pp. 1–4, IEEE.

[14] Abhinau K. Venkataramanan, Cosmin Stejerean, and Alan C. Bovik, "Funque: Fusion of Unified Quality Evaluators," in *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 2022, pp. 2147–2151, IEEE.

[15] Werner Robitza, Rakesh Rao Ramachandra Rao, Steve Göring, and Alexer Raake, "Impact of Spatial and Temporal Information on Video Quality and Compressibility," in *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, June 2021, pp. 65–68, ISSN: 2472-7814.

[16] M.H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[17] Michael King, Zinovi Tauber, and Ze-Nian Li, "A New Energy Function for Segmentation and Compression," in *Multimedia and Expo, 2007 IEEE International Conference on*, Beijing, China, July 2007, pp. 1647–1650, IEEE.

[18] Vignesh V Menon, Christian Feldmann, Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer, "VCA: Video Complexity Analyzer," in *Proceedings of the 13th ACM Multimedia Systems Conference*, June 2022, pp. 259–264.

[19] Ekrem Çetinkaya, Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer, "CTU Depth Decision Algorithms for HEVC: A Survey," *Signal Processing: Image Communication*, vol. 99, pp. 116442, Nov. 2021.

[20] Alexandros Stergiou and Ronald Poppe, "AdaPool: Exponential Adaptive Pooling for Information-Retaining Downsampling," *IEEE Transactions on Image Processing*, vol. 32, pp. 251–266, 2023.

[21] Apple, "HTTP Live Streaming (HLS) Authoring Specification for Apple Devices | Apple Developer Documentation," 2015.

[22] Vignesh V Menon, Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer, "OPTE: Online Per-Title Encoding for Live Video Streaming," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 1865–1869, ISSN: 2379-190X.