Transcoding Quality Prediction for Adaptive Video Streaming

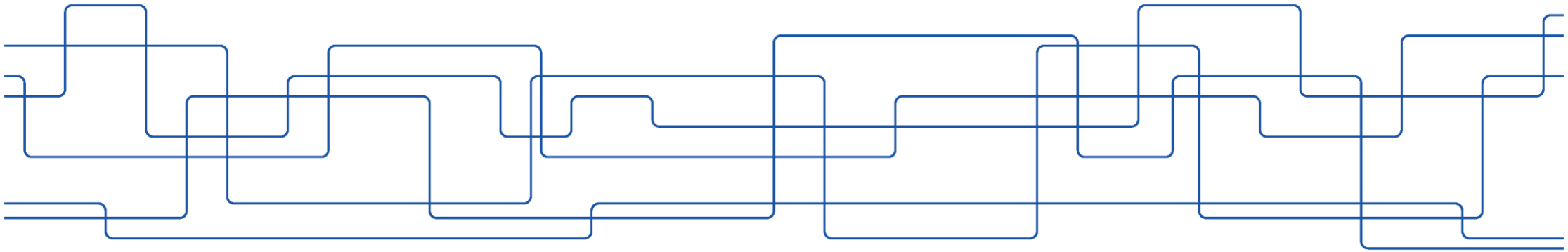
Vignesh V Menon[1], Reza Farahani[1], Prajit T Rajendran[2], Mohammad Ghanbari[3], Hermann Hellwagner[1], Christian Timmerer[1]

[1]Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität, Klagenfurt, Austria

[2]CEA, List, F-91120 Palaiseau, Université Paris-Saclay, France

[3]School of Computer Science and Electronic Engineering, University of Essex, UK

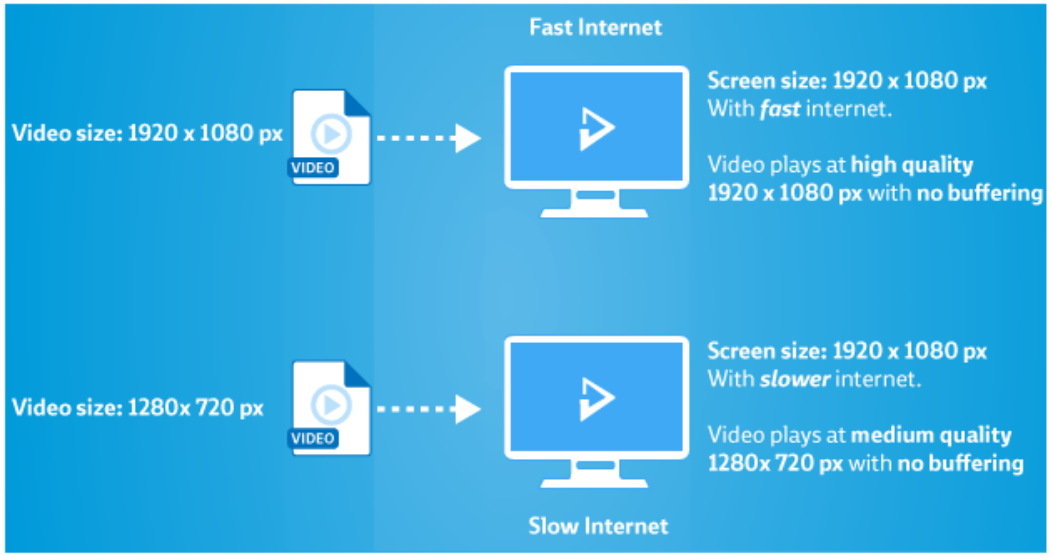
08 May 2023



Outline

Introduction

HTTP Adaptive Streaming (HAS)[1]



Source: https://bitmovin.com/adaptive-streaming/

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.

Motivation

Video transcoding has been considered a prevalent solution for reconstructing video sequences at *in-network servers* (deployed at cloud or edge) in latency-sensitive video streaming applications.

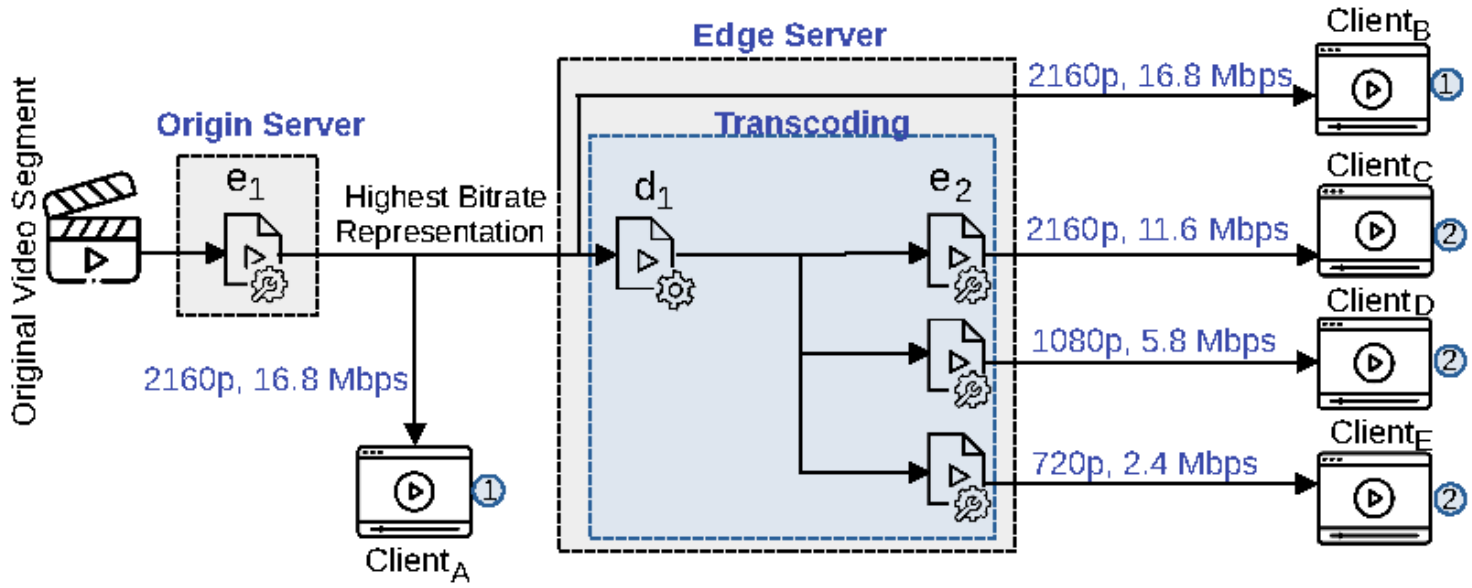


Figure: An example scenario of VQA in adaptive streaming applications. Clients A and B receive the highest bitrate representation of the bitrate ladder, encoded at the origin server (single-stage transcoding). In contrast, Clients C, D, and E receive lower bitrate representations transcoded at the edge server (two-stage transcoding).

Motivation

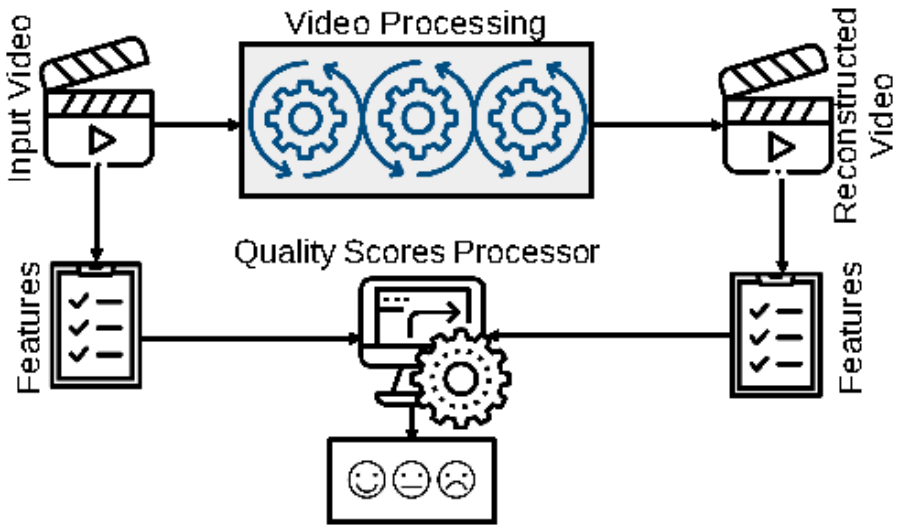


Figure: Workflow of state-of-the-art VQA methods.

VQA is cumbersome in most video streaming applications where:

- The original input video segment is not available as the reference at the destination
- The final reconstructed video segment is not available at the source
- Slow quality-based decision-making is not acceptable for online latency-sensitive services.

M-stage transcoding model

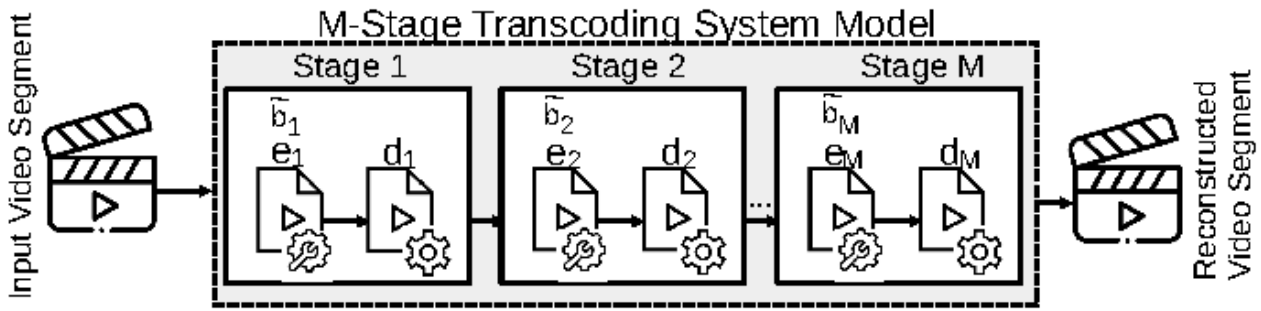
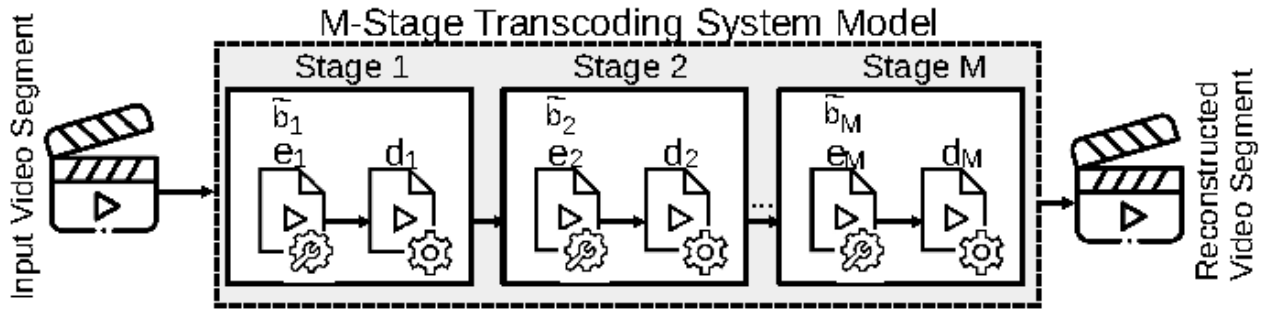


Figure: M-stage transcoding model considered in this paper. Here, e_i and d_i represent the encoding and decoding in i^{th} stage of transcoding, while $\tilde{b}_i$ denotes the target bitrate of e_i where $i \in [1, M]$.

- The generalized M-stage transcoding model for HAS consists of a series of M encoders and M decoders in a chain.
- M=1 transcoding corresponds to the single-stage transcoding while M=2 transcoding corresponds to the two-stage transcoding.

M-stage transcoding model



$$\tau_T = \sum_{i=1}^{M} (\tau_{e_i} + \tau_{d_i}) + 2 \cdot \tau_f \tag{1}$$

VQA at source by predicting the video quality using the input video segment characteristics and the transcoding system characteristics solves the discussed problems.

TQPM Architecture

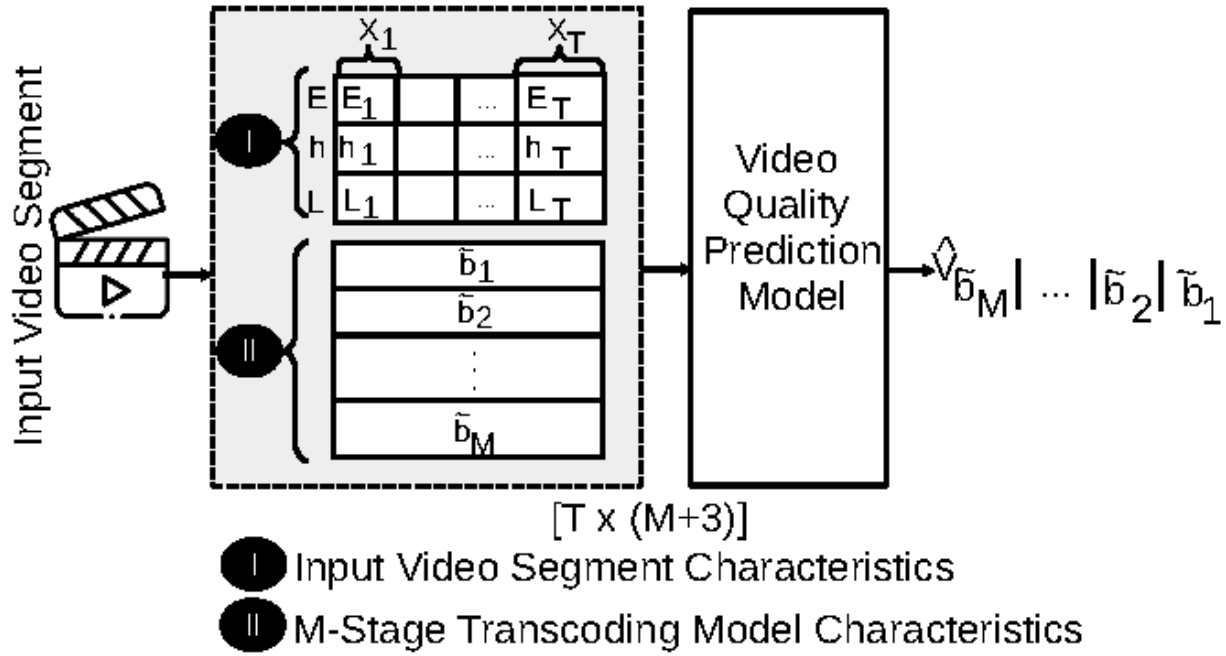


Figure: TQPM architecture

The TQPM architecture comprises three steps:

- input video segment characterization
- transcoding model Characterization
- video quality prediction

Input video segment characterization

Compute texture energy per block

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

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

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.[2,3]

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

[2] 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.

[3] 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.

Input video segment characterization

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

$$h_s = \sum_{k=0}^{K-1} \frac{| H_{s,k}, H_{s-1,k} |}{K \cdot w^2} \tag{4}$$

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

$$L_{s,k} = \sqrt{DCT(0,0)} \tag{5}$$

The block-wise luminescence is averaged per frame denoted as L_s as shown below.[4]

$$L_s = \sum_{k=0}^{K-1} \frac{L_{s,k}}{K \cdot w^2} \tag{6}$$

[4] 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.

Input video segment characterization

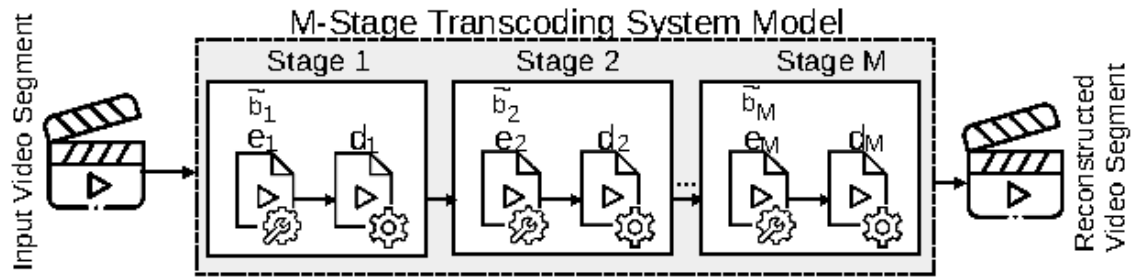
The video segment is divided into T chunks with a fixed number of frames (*i.e.*, f_c) in each chunk. The averages of the E, h, and L features of each chunk are computed to obtain the *reduced reference representation* of the input video segment, expressed as:

$$X = \{x_1, x_2, .., x_T\} \tag{7}$$

where, x_i is the feature set of every i^{th} chunk, represented as :

$$x_i = [E_i, h_i, L_i] \quad \forall i \in [1, T] \tag{8}$$

Phase 2: Transcoding model Characterization



- The settings of the encoders in the M-stage transcoding process, except the target bitrate-resolution pair, are assumed identical.[5]
- The resolutions corresponding to the target bitrates in the bitrate ladder are also assumed to be fixed.

The transcoding model can be characterized as follows:

$$\tilde{B} = [\tilde{b}_1, \tilde{b}_2, .., \tilde{b}_M] \tag{9}$$

where $\tilde{b}_i$ represents the target bitrate of the e_i encoder.

[5] Vignesh V Menon et al. "EMES: Efficient Multi-Encoding Schemes for HEVC-Based Adaptive Bitrate Streaming". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 19.3s (2023). ISSN: 1551-6857. DOI: 10.1145/3575659. URL: https://doi.org/10.1145/3575659.

Phase 3: Video quality prediction

$\tilde{B}$ is appended to x_i, which is determined during the input video segment characterization phase, to obtain:

$$\tilde{x_i} = [x_i|\tilde{B}]^T \quad \forall \tilde{x_i} \in \tilde{X}, \quad i \in [1, T] \tag{10}$$

The predicted quality $\hat{v}_{\tilde{b}_M|..|\tilde{b}_1}$ can be presented as:

$$\hat{v}_{\tilde{b}_M|..|\tilde{b}_1} = f(\tilde{X}) \tag{11}$$

The feature sequences in the series $\tilde{X}$ are input to the LSTM model, which predicts visual quality $\hat{v}$ for the corresponding input video segment and chain of encoders in the transcoding process.

Experimental Setup

Dataset : JVET,[6] MCML,[7] SJTU,[8] Berlin,[9] UVG,[10] BVI[11]
Framerate : 30fps
Encoder : x265 v3.5
Preset : ultrafast

Table: Representations considered in this paper.

Representation ID	01	02	03	04	05	06	07	08	09	10	11	12
r (width in pixels)	360	432	540	540	540	720	720	1080	1080	1440	2160	2160
b (in Mbps)	0.145	0.300	0.600	0.900	1.600	2.400	3.400	4.500	5.800	8.100	11.600	16.800

E, h, L features are extracted using VCAv2.0.

[6] Jill Boyce et al. *JVET-J1010: JVET common test conditions and software reference configurations*. July 2018.

[7] Manri Cheon and Jong-Seok Lee. "Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience". In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.7 (2018), pp. 1467–1480. DOI: 10.1109/TCSVT.2017.2683504.

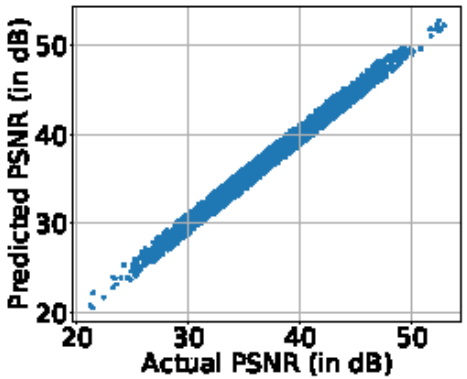
[8] Li Song et al. "The SJTU 4K Video Sequence Dataset". In: *Fifth International Workshop on Quality of Multimedia Experience (QoMEX2013)* (July 2013).

[9] B. Bross et al. "AHG4 Multiformat Berlin Test Sequences". In: *JVET-Q0791*. 2020.

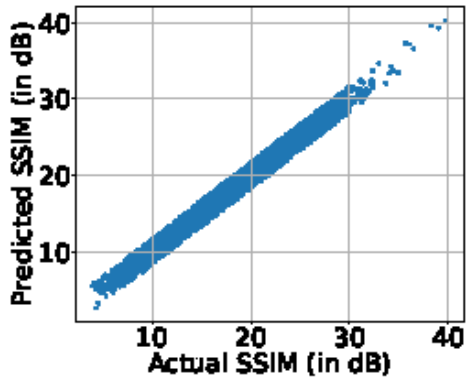
[10] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. "UVG Dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development". In: *Proceedings of the 11th ACM Multimedia Systems Conference*. New York, NY, USA: Association for Computing Machinery, 2020, 297–302. ISBN: 9781450368452. URL: https://doi.org/10.1145/3339825.3394937.

[11] Alex Mackin, Fan Zhang, and David R. Bull. "A study of subjective video quality at various frame rates". In: *2015 IEEE International Conference on Image Processing (ICIP)*. 2015, pp. 3407–3411. DOI: 10.1109/ICIP.2015.7351436.

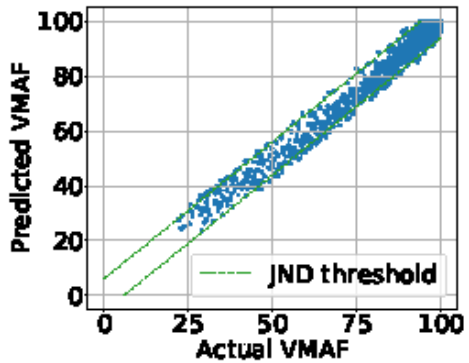
Experimental Results



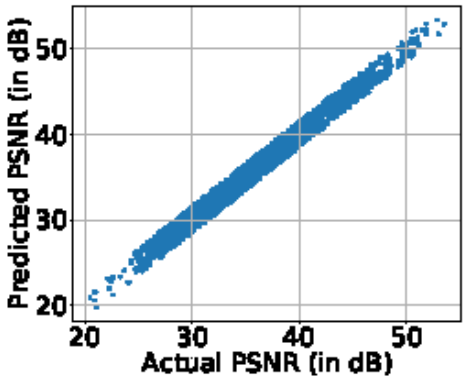
(a)



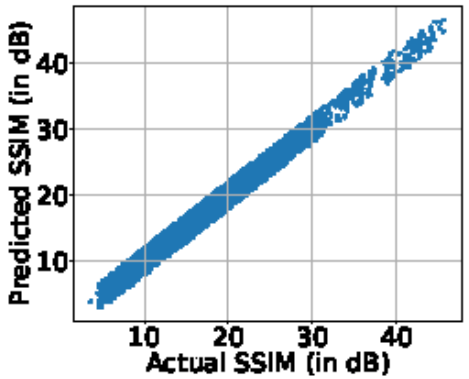
(b)



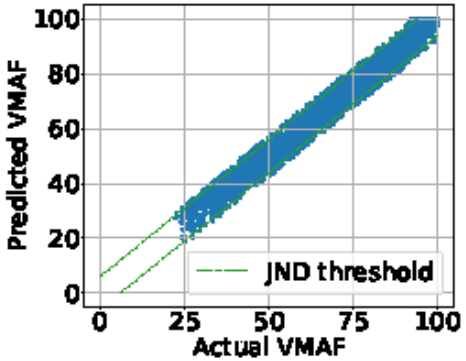
(c)



(d)



(e)



(f)

Figure: Scatterplots of the actual quality and predicted quality for M=1 ((a) PSNR, (b) SSIM, and (c) VMAF, respectively) and M=2 transcoding ((d) PSNR, (e) SSIM, and (f) VMAF, respectively).

Experimental Results

Table: Prediction accuracy of TQPM when M=1 and M=2, respectively, for $\tilde{b}_1$ representations considered in this paper encoded using x265 HEVC encoder.

			PSNR prediction				SSIM prediction				VMAF prediction			
			M=1		M=2		M=1		M=2		M=1		M=2	
$\tilde{b}_1$			R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
b_1	360p	0.145 Mbps	0.82	1.20 dB	-	-	0.89	1.08 dB	-	-	0.87	3.35	-	-
b_2	432p	0.300 Mbps	0.83	1.19 dB	0.84	1.37 dB	0.89	1.14 dB	0.87	1.34 dB	0.87	3.51	0.76	3.38
b_3	540p	0.600 Mbps	0.83	1.19 dB	0.85	1.28 dB	0.88	1.18 dB	0.85	1.21 dB	0.90	4.05	0.84	3.55
b_4	540p	0.900 Mbps	0.83	1.19 dB	0.83	1.22 dB	0.86	1.17 dB	0.86	1.11 dB	0.90	3.83	0.89	3.53
b_5	540p	1.600 Mbps	0.82	1.22 dB	0.82	1.15 dB	0.84	1.19 dB	0.85	1.38 dB	0.90	3.45	0.90	3.44
b_6	720p	2.400 Mbps	0.83	1.26 dB	0.83	1.28 dB	0.82	1.18 dB	0.83	1.57 dB	0.88	2.88	0.91	3.45
b_7	720p	3.400 Mbps	0.81	1.30 dB	0.85	1.23 dB	0.83	1.20 dB	0.82	1.35 dB	0.84	2.89	0.94	3.03
b_8	1080p	4.500 Mbps	0.84	1.28 dB	0.83	1.28 dB	0.88	1.23 dB	0.82	1.34 dB	0.87	2.28	0.95	3.03
b_9	1080p	5.800 Mbps	0.86	1.31 dB	0.87	1.42 dB	0.83	1.29 dB	0.86	1.30 dB	0.87	2.23	0.95	3.34
b_{10}	1440p	8.100 Mbps	0.84	1.39 dB	0.81	1.41 dB	0.87	1.29 dB	0.87	1.32 dB	0.85	2.73	0.96	2.96
b_{11}	2160p	11.600 Mbps	0.79	1.50 dB	0.82	1.31 dB	0.88	1.17 dB	0.84	1.32 dB	0.82	2.58	0.96	3.02
b_{12}	2160p	16.800 Mbps	0.84	1.49 dB	0.79	1.26 dB	0.88	1.19 dB	0.86	1.35 dB	0.86	2.38	0.96	2.99
Average			**0.83**	**1.31 dB**	**0.84**	**1.32 dB**	**0.85**	**1.19 dB**	**0.86**	**1.33 dB**	**0.87**	**3.01**	**0.91**	**3.25**

The average processing time of TQPM for a 4s segment is 0.328s.

Conclusions

- This paper proposed TQPM, an online transcoding quality prediction model for video streaming applications.
- The proposed LSTM-based model uses DCT-energy-based features as *reduced reference* to characterize the input video segment, which is used to predict the visual quality of an M-stage transcoding process.
- The performance of TQPM is validated by the Apple HLS bitrate ladder encoding and transcoding using the x265 open-source HEVC encoder.
- On average, for single-stage transcoding, TQPM predicts PSNR, SSIM, and VMAF with an MAE of 1.31 dB, 1.19 dB, and 3.01, respectively.
- Furthermore, PSNR, SSIM, and VMAF are predicted for two-stage transcoding with an average MAE of 1.32 dB, 1.33 dB, and 3.25, respectively.

Future Directions

- In the future, transcoding between bitrate ladder representations of various codecs shall be investigated.
- Another future direction is defining a decision-making component based on the proposed model in an end-to-end live streaming system.

Q & A

Thank you for your attention!

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

VCA