SMPTE

MOTION IMAGING JOURNAL

Covering Emerging Technologies for the Global Media Community

Machine Learning for Per-Title Encoding

By Daniel Silhavy, Christopher Krauss, Anita Chen, Anh-Tu Nguyen, Christoph Müller, Stefan Arbanowski, Stephan Steglich, and Louay Bassbouss

Introdução:

A motivação para escrever a introdução dos Artigos da SMPTE, não é outra senão convidar os leitores da SET a debruçarem-se sobre artigos científicos que estão em sua maioria na vanguarda das tecnologias que estão sendo desenvolvidas. Dito isso, o convite para ler este presente artigo é outro, é para trazer aos leitores um verdadeiro desafio e uma verdadeira "*open mind*", ao se deparar com um artigo que une aprendizado de maquina e o conceito de encoder por titulo (este introduzido pela NETFLIX em 2016) e, largamente, utilizado na indústria de ***streaming*** desde então.

Não vou mentir, o artigo é desafiador no sentido de trazer uma nova sopa de letrinhas desconhecida do universo ***broadcast*** e muito utilizada no mundo de ***Machine Learning*** e ***Data Miner***, mas eu garanto que você sairá transformado ao final de um artigo que se esforça em mostrar que sempre é possível fazer melhor. Em um mundo cada vez mais raso nos conteúdos das redes sociais, este artigo é o mergulho no profundo conhecimento que tanto faz falta. Boa Leitura!

Tom Jones Moreira

Abstract

Video streaming content varies in terms of complexity and requires title-specific encoding settings to achieve a certain visual quality. Classic "one-size-fits-all" encoding ladders ignore video-specific characteristics and apply the same encoding settings across all video files. In the worst-case scenario, this approach can lead to quality impairments, encoding artifacts, or unnecessarily large media files. A per-title encoding solution has the potential to significantly decrease the storage and delivery costs of video streams while improving the perceptual quality of the video. Conventional per-title encoding solutions typically require a large number of test encodes, resulting in high computational times and costs. In this article, we describe a solution that implements the conventional per-title encoding approach and uses its resulting data for machine learning-based improvements. By applying supervised, multivariate regression algorithms like random forest regression, multilayer perceptron (MLP), and support vector regression, we can predict video quality metric (VMAF) values. These video quality metric values are the foundation for deriving the optimal encoding ladder. As a result, the test encodes are eliminated while preserving the benefits of conventional per-title encoding.

In 2016, Netflix introduced the concept of per-title-encoding. Per-title encoding is based on the fact that different types of video content require different bitrates and encoding settings to achieve a certain quality. In comparison to the conventional one-size-fits-all encoding approach, in which the same, predefined encoding ladder is applied for all types of content, per-title encoding has the potential to significantly decrease the storage and delivery costs of video streams. Easy-to-encode videos such as animations with high redundancy between frames can be delivered with significantly lower bitrates while ideally maintaining or even improving the perceptual quality. In addition, high complexity content like action movies or sport streams, which contain a lot of movement, is streamed with lower resolutions to avoid a lower quality of experience for the viewer.

Keywords

Adaptive bitrate streaming, machine learning, per-title encoding, video multi-method assessment fusion (VMAF)

Introduction

In 2018, 58% of the global application internet traffic stemmed from video streaming applications.[1] Streaming providers, such as Netflix and YouTube, accounted for 15% and 11% of the web traffic, respectively.[1] Netflix states that 70% of its streams end up on connected TVs, while phones, tablets, and PCs account for the remaining 30%.[2] Taking a closer look at connected TVs, more than 100 million 4K UHD TVs were sold in 2018. High-dynamic range (HDR) technology is also becoming a factor in terms of delivering high-quality video content, which is embedded in 60% of 4K UHD sets sold in 2018.[3] In addition, today's media streaming landscape is dominated by adaptive streaming technologies. The main formats utilized in this context are HTTP live streaming (HLS) and dynamic adaptive streaming over HTTP (MPEG-DASH).[4]

These trends resulted in significant increases in cost in terms of content storage and delivery. Content providers are required to support different streaming formats (HLS and DASH) across various platforms (PCs, TVs, and mobile devices). Consequently, video assets need to be encoded, stored, and delivered in multiple qualities and formats. For example, 4K and HDR technology require higher bitrates than 1080p or 720p content.

One way to tackle the high storage and delivery costs is to utilize the latest video codecs, such as VP9, AV1, or H.265. However, many legacy devices do not support these codecs, which is why H.264 is still the dominant codec.[4]

In 2016, Netflix introduced the concept of per-title-encoding.[5] Per-title encoding is based on the fact that different types of video content require different bitrates and encoding settings to achieve a certain quality. In comparison to the conventional one-size-fits-all encoding approach, in which the same, predefined encoding ladder is applied for all types of content, per-title encoding has the potential to significantly decrease the storage and delivery costs of video streams. Easy-to-encode videos such as animations with high redundancy between frames can be delivered with significantly lower bitrates while ideally maintaining or even improving the perceptual quality. In addition, high complexity content like action movies or sport streams, which contain a lot of movement, is streamed with lower resolutions to avoid a lower quality of experience for the viewer.

Within the per-title-encoding process, the optimal encoding settings of a video clip are identified in the complexity analysis step. The most common approach for determining the complexity of a video is to run multiple test encodes. Based on these test encodes, a content-specific bitrate/resolution ladder is derived and applied to the entire video clip.

The major downside of this per-title-encoding approach comes with the fact that it is computationally very heavy and typically requires a large number of test encodes to derive a sufficient amount of data. While such an approach is affordable for large companies with ample financial and computational resources, it becomes a downside for smaller companies. The latter would need to carefully evaluate whether the storage and delivery gains outweigh the computational costs.

The purpose of our work is to avoid the cost-intensive test encodes, while preserving the benefits of the traditional per-title encoding approach. Rather than performing test encodes, we provide a solution that predicts the quality of a video based on a given set of encoding settings by using machine learning techniques such as neural networks and random forest. This way, we can identify the optimal encoding ladder while reducing processing time and costs needed for generating the test encodes.

Related Work

De Cock et al.[5] described a method for encoding per-title video-on-demand (VoD) content. Based on a complexity analysis, bitrate-resolution pairs positioned closely to the convex hull are derived. They further improve this approach by applying a chunk-based multipass encoding process. As a result, title- and chunk-based encoding approaches outperform conventional approaches in terms of storage savings and video quality.

Takeuchi et al.[6] estimated bitrate-quality curves using the just noticeable difference (JND) scale. Based on a feature set that includes the quantization parameter (QP), resolution, and video quality metrics such as the peak-signal-to-noise ratio (PSNR), video multi-method assessment fusion (VMAF), and structural similarity index (SSIM), a support vector regression (SVR) model is applied to estimate the JND scores. Their results show that a JND-based encoding ladder results in smaller storage sizes compared to conventional encoding ladders.

Chen et al.[7] used a probability distribution of viewport and bandwidth to minimize the streaming cost while maintaining high streaming quality. Their A/B testing results demonstrated bandwidth savings of 9.7% without degrading the viewer's quality of experience.

Rassool[8] examined the correlation between the subjective mean opinion score and the computed VMAF score. The results indicate that a VMAF score of 93 or higher is sufficient to produce a video that is either indistinguishable from the original or with noticeable, but not annoying distortion.

Per-Title Encoding

In general, the efficiency of a video codec is correlated with the input video's spatial and temporal redundancy. A video that contains several movements and scene changes is more difficult to encode than a video where most parts are redundant or slowly changing over time. We analyzed 60 movie trailers (with a duration of 3 min and a resolution of 1080p) and plotted their resulting bitrate/resolution value pairs in **Fig. 1**.

We found that animated trailers achieve Y-PSNR values of 40–45 dB at bitrates ranging from 0.7 to 1.3 Mbit/s. More complex content like action trailers require a much higher bitrate of 2.5–7.5 Mbit/s to achieve similar dB values of 40–45. The conclusion of this is obvious: diverse content types require different bitrate settings to achieve a certain quality. Therefore, applying a per-title encoding-ladder approach

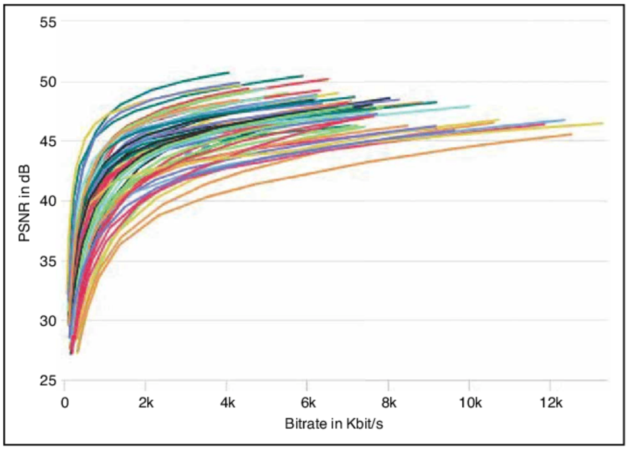

FIGURE 1. PSNR/bitrate pairs of 60 movie trailers at a resolution of 1080p. The movies differ in terms of complexity and therefore require different bitrates to achieve a certain quality.

introduces major advantages compared to the conventional approach of a "one-size-fits-all" encoding ladder.

In a per-title encoding solution, low-complexity content is encoded using significantly lower bitrates and therefore requires less storage space while saving delivery costs. In addition, low-complexity videos can be delivered at a high resolution using a small bitrate. Consequently, the perceived quality for the user is significantly increased at lower bitrates. Note that, in an adaptive streaming solution, the delivered quality not only depends on the encoding settings, but also on the media player's adaptation logic and client-side network conditions.

Table 1. Classic encoding ladder for content with medium complexity.

Resolution	Bitrate (kbit/s)	VMAF	PSNR (dB)
PSNR (dB)	320	30.87	29.8
384 × 288	400	40.2	30.4
512 × 384	750	55.8	31.62
640 × 480	1200	67.5	32.89
720 × 480	1900	72	33.4
1280 × 720	3000	86.2	37.43
1280 × 720	4500	88.1	38
1920 × 1080	6000	94.01	43.12
1920 × 1080	7800	95.2	44.6

Table 2. Title-based encoding ladder for content with medium complexity.

Resolution	Bitrate (kbit/s)	VMAF	PSNR (dB)
640 × 480	320	49.5	31.4
1280 × 720	750	68.5	31.4
1280 × 720	1200	76.2	35.3
1280 × 720	1700	82.3	37.4
1920 × 1080	3000	88.5	39.3
1920 × 1080	5000	93.2	42.5

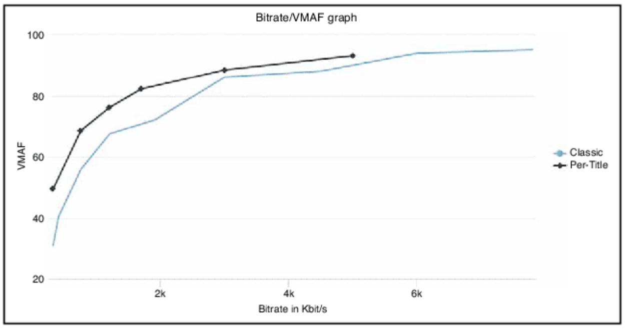

FIGURE 2. Resulting interpolated VMAF values from conventional and per-title encoding ladders. The per-title solution provides higher quality scores by utilizing the ideal resolution for a given bitrate.

Classic Versus Per-Title Encoding Ladders

The aforementioned conclusions are illustrated in **Tables 1** and **2** and **Fig. 2**. A video with medium complexity is encoded in H.264 using the conventional and per-title encoding ladder methods. The VMAF metric, also developed by Netflix, was employed to determine the quality of each encode. Compared to PSNR, VMAF offers a better accuracy in measuring the human perception of video quality and provides consistent results across various content. The first thing to notice is that the per-title ladder contains fewer bitrate/resolution pairs than the conventional encoding ladder. This arises for various reasons: the lowest bitrate of 320 kbit/s delivers its optimal quality (highest VMAF score) at a resolution of 480p. Hence, all lower resolutions can be omitted. Additionally, JND plays a crucial role when identifying optimal bitrate/resolution pairs. JND is defined as the value by which something must be changed for a difference to be noticeable. Since six VMAF points equal to one JND, each encode should be at least six VMAF points apart.[9] However, this is not the case for the 3 and 4.5 Mbit/s representations or the 6 and 7.8 Mbit/s representations in the conventional encoding ladder. As a result, the video is stored in qualities that show no perceivable visual difference.

Furthermore, a VMAF score of 93 is sufficient in producing a video that is either indistinguishable from the original or with noticeable but not annoying distortion.[8] While the title-based encoding ladder is capped at approximately 93 VMAF points, the conventional ladder goes up to a score of 95.2.

Table 3 summarizes the results of both encoding approaches for a streaming session with 20 Mbit/s available on the client side. Compared to the conventional encoding ladder approach, storage costs are 52% lower for per-title solutions. Since the client consistently delivers the best quality, the network traffic of the per-title variant is 36%

Table 3. Evaluation of a conventional- and a per-title encoding ladder for a medium-complexity content with a duration of 10 minutes. The per-title solution saves 52% in storage costs and 36% in delivery costs while delivering an average level of video quality.

	Average values			
	Bitrate (kbit/s)	VMAF	PSNR (dB)	Storage (MB)
Conventional	7648.18	94.92	44.37	1397.7
Per-Title	4941.75	93.06	42.41	675.2
Difference Abs	+2706.43	−1.86	−1.96	675.2
Difference Perc	+36%	−1%	−4%	+52%

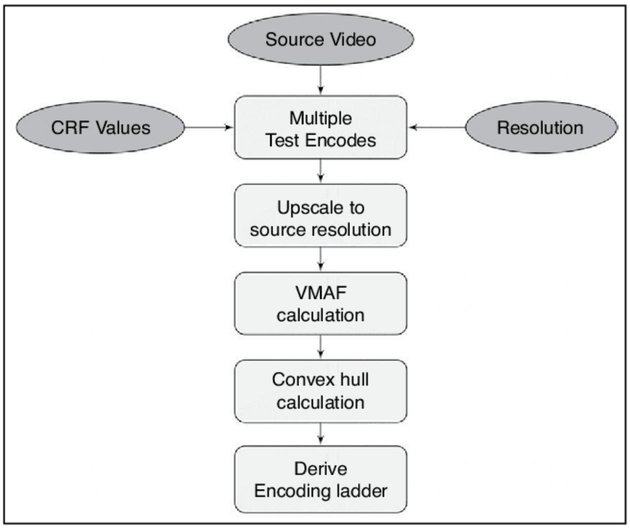

FIGURE 3. Basic per-title encoding workflow.

lower than that of the conventional method. Although the conventional solution delivers higher VMAF and PSNR values, the perceived quality for the viewer is similar due to VMAF scores of 93+ in both cases. Thus, the loss in VMAF is of no consequence; instead, the content provider benefits from massive bitrate and network savings while delivering optimal video quality.

Determining the Per-Title Encoding Ladder

The conventional workflow in determining a per-title encoding ladder is depicted in **Fig. 3**. The source video, target encoding settings (resolution, codec, and GOP size), and a list of constant rate factor (CRF) values serve as the input for multiple test encodes.

To determine the video quality of the resulting test encodes, metrics such as PSNR, VMAF, and SSIM are required. In the context of this article, we use the default VMAF model, which predicts the quality of videos displayed on a 1080p HDTV in a living-room-like environment. However, the previously described principles are also valid for other video quality metrics. By definition, the VMAF values can only be calculated on videos that have the same resolution. Therefore, prior to the actual VMAF calculation, the encoded videos are upscaled to the resolution of the source video. To match the VMAF model, we exclusively deployed 1080p source videos in this article. All test encodes, except for the 1080p outputs, are then upscaled to this resolution using the bicubic upscaling filter. The bitrate/VMAF values of the different resolutions form a boundary called the convex hull. The per-title encoding ladder is derived from the selection of bitrate/resolution pairs positioned closest to the convex hull.[10] An example of a convex hull is depicted in **Fig. 4**.

An important aspect of this process is the fact that only a limited number of data points can be generated from the test encodes. Therefore, interpolation and extrapolation are required to derive the bitrate/quality curves, as shown in **Fig. 4**. Limiting the number of test encodes may lead to an insufficient number of data points and a distorted representation of the quality curves.

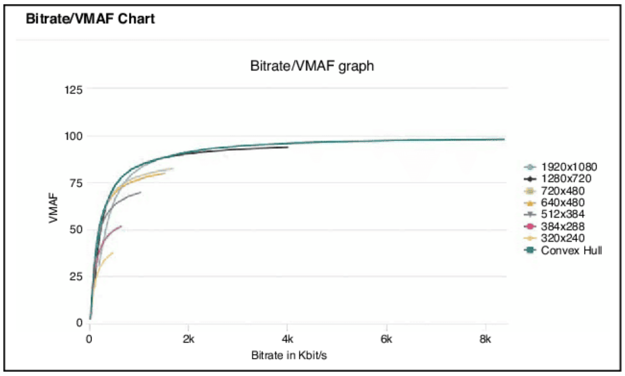

FIGURE 4. VMAF/bitrate values and the resulting convex hull derived from multiple test encodes of different resolutions.

Since the complexity of a video may vary throughout its duration, it is reasonable to extend per-title encoding to scene-based encoding. Furthermore, the described per-title encoding approach can be optimized by only test-encoding the high-complexity parts of the source video.[10] Both optimizations are ongoing research and not considered in this article. Instead, the focus is on improving the computational time of the brute-force approach by applying machine learning models to predict the VMAF values, thus omitting the need for test encodes.

Per-Title Encoding Using Machine Learning

As described in the previous section, the optimal encoding ladder can be derived by applying a large number of test encodes (typically 7–12 per resolution), determining multiple bitrate/VMAF pairs for each resolution, interpolating and extrapolating the data points, and calculating the convex hull. An algorithm that predicts the VMAF values and reduces the high number of test encodes at the same time offers significant advantages in terms of computational costs and scalability.

Prediction algorithms are typically based on machine learning techniques. The task of predicting VMAF scores based on predefined bitrates can be categorized as a supervised learning multivariate regression task (due to the use of labeled reference data). **Figure 5** depicts a high-level overview of this approach. The training data consists of test encodes, as well as corresponding video metadata and additional information (e.g., number of scene changes). The algorithm recognizes the dependencies between these variables and the targeted quality metrics and applies this knowledge to future videos. In the prediction phase, only the video metadata and a predefined number of target bitrates are fed into the model. The model's outputs are the corresponding VMAF values, which can then be used to derive the optimal encoding settings. To support the video analysis and metadata extraction, further machine learning algorithms (especially for image processing) can be applied.

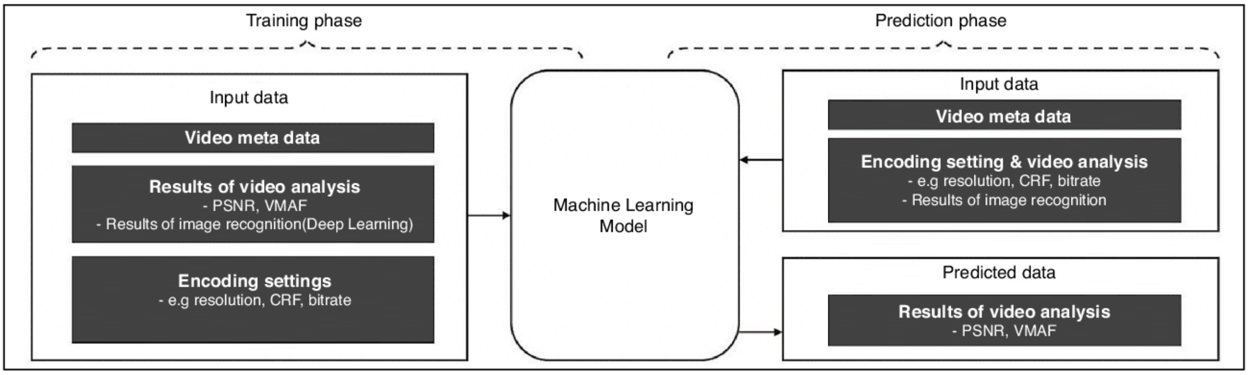
Training phase
Input data
Video meta data
Results of video analysis
- PSNR, VMAF
- Results of image recognition(Deep Learning)
Encoding settings
- e.g resolution, CRF, bitrate
Machine Learning Model
Prediction phase
Input data
Video meta data
Encoding setting & video analysis
- e.g resolution, CRF, bitrate
- Results of image recognition
Predicted data
Results of video analysis
- PSNR, VMAF

FIGURE 5. Simplified workflow of machine learning-based per-title encoding.

To continuously refine and improve the process, additional test encodes of the input video are carried out in the background. As this is not a time-critical calculation, a small, inexpensive machine can be used for such purposes. The comparison between the machine learning predicted VMAF values and the actual results of the various encodes provides vital information about the system's accuracy.

Methodology and Results

The complete process of training and evaluating our machine learning models is divided into five parts.

Data Preprocessing

In the initial analyzation and preprocessing step, the raw data is transformed into an understandable format. This step plays a significant role in the overall workflow due to incomplete and inconsistent data errors that are usually discovered within raw data. In this step, two sets of raw data are available ("video" and "encode").

The "video" dataset provides general information and metadata of our source videos, which are movie trailers with a duration of 120–150 seconds. The information included in the video dataset are duration, size, bitrate, video codec, width, height, and framerate of each video. The "encode" dataset specifies information about the respective test encodes for each of the source videos. Each source video in the reference dataset was encoded with resolution, CRF, and codec settings, as depicted in **Table 4**. In our tests, we exclusively used the H.264 codec. However, this approach is codec-agnostic and not limited to H.264. Seven target resolutions with 12 CRF values lead to a total of 84 test encodes for a single asset.

After data cleansing, the "encode" and "video" datasets were merged for further preprocessing, resulting in a total of 11,011 video samples.

Features Engineering

As part of the feature engineering step, we generated two additional attributes through feature combination. The "resolution" attribute was generated by combining the width and height columns (e.g., 1920 × 1080). Each combined pair was assigned with a numerical label and labeled as "*res_encode*." In addition, a new attribute, "size_mean," was computed by dividing the size of each clip by the product of duration, width, and height. To normalize the data, each frame rate was assigned with a label "encode" (similar to the generated "resolution" attribute) and renamed as "*FPS_encode*." Lastly, the "*bitrate_video*" attribute values were scaled for consistency purposes by dividing all values by 10^6.

Prior to applying the model, we computed the correlation between each variable (*CRF*, *res_encode*, *bitrate_video*, *FPS_encode*, and *size_mean*) and *VMAF* to determine the independent and dependent variables that will serve as input for the machine learning pipeline. As shown in **Fig. 6**, we found that the CRF, bitrate video, resolution, and frame rate attributes showed the highest correlation with VMAF, and therefore selected them as our primary input.

Evaluation Criteria

To predict the VMAF scores for specific bitrates, we focused on the following three supervised machine learning models, which are further explained in the following section: (1) SVR, (2) random forest, and (3) multilayer perceptron (MLP). We fivefold cross-validated the 11,011 sample clips to select the appropriate machine learning models. After cross-validation, we split the data between training and testing subsets (80/20) and determined the root mean square error (RMSE) and the R-squared values.

R-squared is a statistical measure of fit that indicates how much variation of a dependent variable is defined by the independent variable(s) in a regression model. It can be calculated with the equations depicted below, with y_i being the actual value, $\bar{y}$ the mean of the observed data, and f_i as the predicted value for observation i

$$\text{Explained Variation} = \sum_{1}^{n} (y_i - f_i)^2 \tag{1}$$

$$\text{Total Variation} = \sum_{1}^{n} (y_i - \bar{y})^2 \tag{2}$$

$$R^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}. \tag{3}$$

The RMSE equation quantifies the extent to which the predicted value for an observation matches its corresponding true value. The RMSE is calculated by

$$\mathrm{RMSE} = \sqrt{\tfrac{1}{n}}\sum\nolimits_{1}^{n} (y_i - f_i)^2. \quad (4)$$

Model Selection and Evaluation

The error values for each of the machine learning models are presented in **Table 5**.

The SVR model is a binary linear classifier. It uses the same principles as the support vector machine (SVM) method albeit with a few minor differences. Most importantly, a margin of tolerance (ε) is fixed, to fit the error within a certain threshold. Based on our variables, the aim was to select the optimal non-negative tuning parameter (C) within a range of 1–5. The other hyperparameters were set to "default." Results showed that the optimal value predicted for C was 4, with an R-squared value of 0.942, and an RMSE of 6.64.

The random forest regressor (RFR) is an ensemble learning method that uses a series of decision trees. It makes predictions by combining decisions from a sequence of base models, which are independently constructed using different subsamples of the available training data. This method was tested based on hyperparameter tuning, in which our code would cease to run once the optimal R-squared value was achieved. The tree depth was set within a range of 2–20, and the number of trees computed was 10, 50, and 100. We found that the optimal tree depth and the number of trees are 14 and 100, respectively, with an R-squared value of 0.946. The RFR model had the lowest RMSE with a value of 3.22.

MLP is a type of neural network that consists of three main layers: (1) input, (2) hidden layer, and (3) output. For this particular model, the hidden layer was used to reduce the computational time. This hidden layer consisted of a neuron interval of [3, 8, 12, 15, 20, 25, 30, 40, 45, 50]. The highest R-squared value of 0.937 was obtained at hidden layer neuron 12 and had an RMSE of 5.82.

Model Integration

The RFR produced the best results in terms of its RMSE and R-squared values. Therefore, we added the RFR model to our existing solution. **Tables 6–8**

Table 4. Settings for the test encodes of a video. Seven target resolutions with 12 CRF values lead to a total of 84 test encodes for a single asset.

Codec	Resolutions	Resolutions
H.264	1920 × 1080, 1280 × 720, 720 × 480, 640 × 480, 512 × 384, 382 × 288, 320 × 240	18, 19, 20, 22, 25, 27, 30, 35, 40, 45, 50, 55

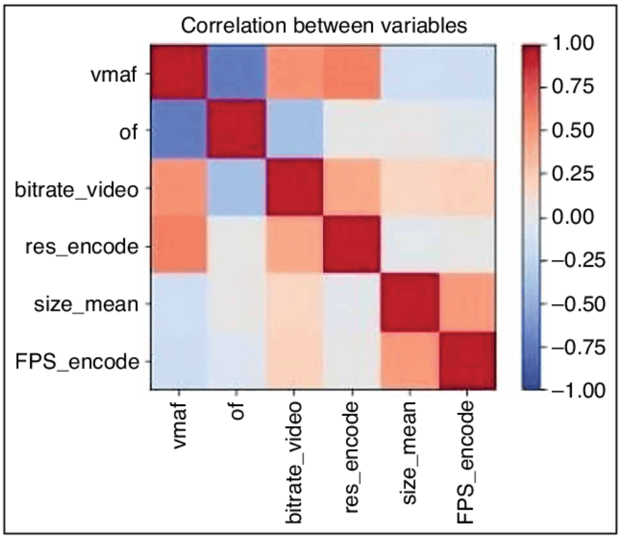

FIGURE 6. Correlation between the different variables. CRF, bitrate video, resolution, and frame rate attributes show the highest correlation with VMAF.

Table 5. Evaluation of the different machine learning models in terms of R-squared values and RMSE.

Model	RMSE	R-squared
RFR	3.22	0.946
MLP	5.82	0.937
SVR	6.64	0.942

Table 6. Classic encoding ladder for a movie trailer.

Resolution	Bitrate (kbit/s)	VMAF
320 × 240	320	60.3
384 × 288	400	67.8
384 × 288	750	78.9
640 × 480	1200	84.7
720 × 480	1900	88.9
1280 × 720	3000	92.2
1280 × 720	4500	94.1
1920 × 1080	6000	94.9
1920 × 1080	7800	96

Table 7. Per-title encoding ladder for a movie trailer.

Resolution	Bitrate (kbit/s)	VMAF
720 × 480	320	71.0
720 × 480	620	80.2
1280 × 720	1220	86.5
1280 × 720	3220	93.1

Table 8. Predicted per-title encoding ladder for a movie trailer.

Resolution	Bitrate (kbit/s)	VMAF	Predicted VMAF
640 × 480	320	70.8	70.1
1280 × 720	620	78.8	80.7
1280 × 720	870	83.3	86.8
1280 × 720	1820	89.3	93.0

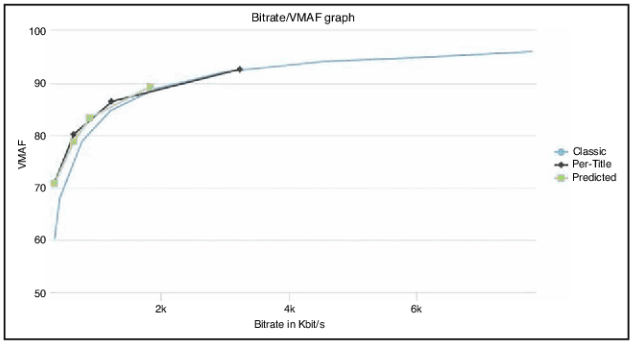

FIGURE 7. Evaluation of different encoding ladders. The per-title encoding ladder derived from multiple test encodes produces the best results in terms of bitrate/quality pairs. The predicted ladder achieves similar results and outperforms the conventional encoding ladder.

Table 9. Evaluation of a conventional-, a predicted per-title-, and a standard per-title encoding ladder for a movie trailer. The per-title solutions offer significant storage and bandwidth savings while delivering approximately the same quality like the conventional encoding ladder.

	Classic	Per-Title	Predicted-Per-Title
Bitrate kbit/s	7618	3165	1775
Bitrate Difference		58.4	76.7
VMAF	95.1	92.1	88.7
VMAF Difference		−3.1%	−6,7%
Storage MB	461.5	93.5	63.1
Storage Difference		79.7%	86.3%

illustrate the resulting encoding ladders for a single movie trailer derived from the conventional "one-size-fits-all" encoding approach, the per-title encoding approach (described in the previous section), and a predicted-per-title encoding ladder using the RFR model, respectively.

In the conventional encoding ladder, certain bitrate/resolution pairs such as 720p@4.5 Mbit/s and 1080p@6 Mbit/s are less than six VMAF points (1 JND) apart. Additionally, the conventional encoding ladder produces qualities with VMAF scores that are higher than 93. In comparison, the per-title and predicted-per-title ladders stay below a VMAF score of 93 and resulted in qualities with at least six VMAF points apart. The predicted ladder peaks at a bitrate of 1820 kbit/s, which results in a VMAF score of 89.3. The predicted VMAF score was higher than the actual value, which is why a real VMAF score of 93 was not obtained (refer to **Table 8**).

Figure 7 shows the interpolated VMAF distribution for different bitrates. The per-title and predicted-per-title encoding ladders deliver better video quality than the conventional encoding ladder by assigning the optimal resolution for each of the target bitrates.

Table 9 summarizes the results of all three encoding approaches for a streaming session with 20 Mbit/s available on the client side. In comparison to the conventional encoding ladder, both per-title solutions save storage costs, by 79.7% and 86.3%. In addition, the average bitrate is up to 76.7% lower for the per-title solutions. Although the conventional solution delivers a higher VMAF value, the per-title encoding solutions produce a similar quality of experience. The standard per-title encoding approach delivers an average VMAF score of 92.1, while the predicted-per-title encoding ladder achieves 88.7 VMAF points.

Conclusion

In this article, we compared the conventional "one-size-fits-all" to per-title encoding approaches. We illustrated how per-title solutions can significantly decrease the storage and delivery costs of video streams while maintaining and even improving the perceived quality for the viewer at the same time. To avoid the necessary, computationally heavy complexity analysis of a standard per-title encoding approach, we proposed the use of machine learning techniques. In this context, we evaluated the use of different types of supervised multivariate regression algorithms to predict the required quality metric scores. The RFR showed the best results in terms of its RMSE and R-squared values. Integration of the RFR into our existing workflow showed that title-specific encoding ladders based on VMAF predictions can outperform conventional encoding ladders. However, the accuracy of the prediction greatly depends on the characteristics of the input video. Thus, a large sample size and a provider or content-specific customized model is required to achieve the optimal results.

Future work includes further refinement of the existing models by providing a larger set of reference content. Moreover, our approach can be improved by restricting the complexity analysis to only the complex parts of a movie and applying the per-scene encoding method. Image feature extraction and classification (based on neural network algorithms) can also be utilized to provide additional metadata for the machine learning models.

References

1. Sandvine, “The Global Internet Phenomena Report,” 2018. [Online]. Available: https://www.sandvine.com/hubfs/downloads/phenomena/2018-phenomena-report.pdf
2. P. Kafka, “You Can Watch Netflix on Any Screen You Want, But You’re Probably Watching It on a TV,” Mar. 2018. [Online]. Available: https://www.recode.net/2018/3/7/17094610/netflix-70-percent-tv-viewing-statistics
3. P. Kurz, “Report: 100M 4K UHD Sets to be Sold Worldwide in 2018,” May 2018. [Online]. Available: https://www.tvtechnology.com/news/report-100m-4k-uhd-sets-to-be-sold-worldwide-in-2018
4. S. Lederer, “Video Developer Report 2018,” 2018. [Online]. Available: http://go.bitmovin.com/download-the-bitmovin-2018-video-developer-report
5. J. De Cock, Z. Li, M. Manohara, and A. Aaron, “Complexity-Based Consistent-Quality Encoding in the Cloud,” *Proc. 2016 IEEE Int. Conf. Image Process. (ICIP)*, pp. 1484–1488, Sep. 2016.
6. M. Takeuchi, S. Saika, and Y. Sakamoto, “Perceptual Quality Driven Adaptive Video Coding Using JND Estimation,” *Picture Coding Symp. (PCS)*, pp. 179–183, Jun. 2018.
7. C. Chen, Y.-C. Lin, S. Benting, and Anil Kokaram, “Optimized Transcoding for Large Scale Adaptive Streaming Using Playback Statistics,” *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, pp. 3269–3273, Oct. 2018.
8. R. Rassool, “Vmaf Reproducibility: Validating a Perceptual Practical Video Quality Metric,” *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, pp. 1–2, Jun. 2017.
9. J. Ozer, “Finding the Just Noticeable Difference With Netflix VMAF,” Sep. 2017. [Online]. Available: https://www.linkedin.com/pulse/finding-just-noticeable-difference-netflix-vmaf-jan-ozer/.
10. A. Aaron, Z. Li, M. Manohara, J. De Cock, and D. Ronca, “Per-Title Encode Optimization,” 2015. [Online]. Available: https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2

About the Authors

Daniel Silhavy is a scientist and project manager at the Business Unit Future Applications and Media (FAME) of the Fraunhofer Institute for Open Communication Systems (FOKUS) (Fraunhofer FOKUS, FAME), Berlin, Germany. He is also the lead developer of the dash.js project and the development coordinator for the 5G-Media Action Group (5G-MAG) Reference Tools. His main area of expertise includes adaptive bitrate (ABR) streaming focusing on MPEG-DASH, digital rights management (DRM), 5G media streaming, and video-encoding. He actively contributes to standardization work done in DASH-IF, 5G-MAG, and Consumer Technology Association-Web Application Video System (CTA-WAVE). Silhavy has published various technical papers covering topics around ABR media delivery and streaming analytics.

Christopher Krauss received a PhD degree (Dr.-Ing.) from the Technical University of Berlin (TU Berlin) with his dissertation “Time-Dependent Recommender Systems for the Prediction of Appropriate Learning Objects.” He works as media & data science lead at Fraunhofer FOKUS, FAME. He specializes in the research and development of topics dealing with machine learning, data science, technology enhanced learning, and connected TVs, and has been involved in multiple public funded projects (e.g., mEDUator, TripleAdapt, EXPAND+ERWB^3, Smart Learning I & II, Digi-Hand, FI-Content I & II, User Centric Networking, and Global ITV) and managed many projects for different national and international industry customers.

Anita Chen has over ten years of technical management in the U.S., Germany, and China, with a focus on the research and development, automotive, health, and media industries. Chen received a BS degree from Business Information Technology, Virginia Tech, Blacksburg, VA, in 2010, and an MS degree from Information Technology Management, TU Berlin, in 2020.

Anh-Tu Nguyen is a researcher and software developer at Fraunhofer FOKUS, FAME. He received bachelor’s and master’s degrees in industrial engineering from TU Berlin. Currently, he is involved in the deep encode project that optimizes video streaming processes using machine learning techniques. He is interested in the applications of machine learning in the computer vision field.

Christoph Müller studied computer science at TU Berlin. He received a degree from FOKUS, with the completion of his thesis “Machine Learning-Supported Adaptive Streaming Analytics” in 2018. He is currently employed as a project manager at FAME, where he is the lead technical project manager for “Deep Encode,” an AI-based solution for per-title and per-scene video encoding. His area of expertise lies in the research and development of modern web applications and machine learning in the context of video encoding, adaptive media streaming and streaming analytics, and has published numerous technical papers on these topics.

Stefan Arbanowski is director of Fraunhofer FOKUS, FAME. Arbanowski received PhD and MSc degrees in computer science from TU Berlin. Currently, he is managing Fraunhofer FOKUS' NewTV activities, bundling expertise in the areas of interactive applications, web technologies, streaming and connected TV-channeling those toward networked media environments featuring live, on demand, context-aware, and personalized interactive content.

Stephan Steglich is director of the business unit at Fraunhofer, FOKUS, FAME. Steglich received MSc and PhD degrees in computer science from TU Berlin in 1998 and 2003, respectively. He is managing international and national level research activities and has been an organizer and a member of program committees of several international conferences. He is actively involved in standardization activities in these research areas and gives lectures at TU Berlin.

Louay Bassbouss is a scientist and senior project manager research and development at Fraunhofer, FOKUS, FAME. He works on future web applications, multiscreen technologies and standards, and 360° video technologies. Bassbouss actively participates in various standardization groups in the World Wide Web Consortium (W3C), Hybrid broadcast broadband TV (HbbTV), and CTA. He is the co-chair of the W3C Second Screen Working and Community Groups and actively contributes to various testing activities in HbbTV and CTA-WAVE.

CONNECT WITH SMPTE FOR A CHANGE

Together, we are spearheading award winning online and in person education across the globe with our preeminent expert members.

SCAN ME

SMPTE