Forward-looking content aware encoding for next generation UHD HDR WCG HFR

Thomas Guionnet, Mickaël Raulet, Thomas Burnichon
ATEME
Rennes, France
t.guionnet@ateme.com, m.raulet@ateme.com, t.burnichon@ateme.com

Abstract - *This paper shows how the future challenges of video compression can be met by leveraging both expertise in optimized compression and latest advances in artificial intelligence, thus spending exactly the right amount of bits on each piece of video information. As we envision a world of life-like TV, enhanced in all possible degrees of freedom, immersive 8K HDR WCG HFR is just knocking at our door. Although technology is improving at a constant pace, processing and transmitting such an amazing amount of information remains an incredible challenge. The BBC transmitted a UHD HDR live feed of the FIFA 2018 World Cup at 36 Mbps. The ATEME TITAN Live encoder does it already at 22Mbps and lower. Even though the future ITU-T ISO/IEC VVC is promising a 50% coding efficiency gain compared to HEVC, compression will not be enough, thus leading to strategies like UHD forum Phase B guidelines recommended content aware encoding [1].*

INTRODUCTION

It is well known how internet media consumption grows continuously [2]. Regarding specifically the TV industry, with major new OTT services launch expected in 2019 by industry giants like Disney, AT&T and Apple, the OTT market is continuing to effectively disrupt the traditional TV market. OTT is now the preferred medium for very large audiences, including for sports events. At the same time, the market is very polarized, content costs are increasing, so the barrier for entry is not as small as one might think, and profitability remains a challenge. Simultaneously, video assets quality is continuously increasing, with 4k and HDR/WCG progressively becoming mainstream while 8k and HFR are under development. This is where efficient video compression SW can contribute by reducing distribution costs while also maximizing QoE. The BBC transmitted a UHD HDR live feed of the FIFA 2018 World Cup at 36 Mbps. ATEME TITAN Live encoder does it already at 22Mbps and lower. The future ITU-T ISO/IEC VVC is promising a 50% coding efficiency gain compared to HEVC. Still, strategies like UHD forum Phase B guidelines recommended content aware encoding [1] can significantly increase bandwidth savings, especially when relying on Artificial Intelligence (AI).

Apple recommends fixed OTT profiles ladders for HLS authoring [3]. These ladders represent average sequence characteristics. By adapting the OTT profiles distribution for each content, one can take advantage of lower complexity content to save bitrate or identify complex contents to improve their visual quality. In the examples of Table 1, and Table 2, the result of content adaptation for two movies contents are reported against Apple recommendations. These contents are simple enough to significantly lower both the profiles bitrates and the number of profiles. Thus, content aware encoding, or content adaptive streaming, saves not only bandwidth, but also storage. One also notes that the bitrate distribution amongst resolutions is different from the recommendation. Indeed, spatial characteristics of the content can change with resolution.

HLS		ATEME content adaptive	
Resolution	Bitrate (kbps)	Resolution	Bitrate (kbps)
1920 x 1080	7800	-	-
1920 x 1080	6000	-	-
1280 x 720	4500	-	-
1280 x 720	3000	1920 x 1080	3269
960 x 540	2000	1536 x 864	1892
768 x 432	1100	1024 x 576	895
768 x 432	730	832 x 468	530
640 x 360	365	640 x 360	289
416 x 234	145	-	-

TABLE 1: AN EXAMPLE OF CONTENT ADAPTED OTT PROFILES COMPARED TO HLS RECOMMENDATION FOR H.264/AVC.

HLS		ATEME content adaptive	
Resolution	Bitrate (kbps)	Resolution	Bitrate (kbps)
3840 x 2160	16800	-	-
3840 x 2160	11600	-	-
2560 x 1440	8100	3840 x 2160	8687
1920 x 1080	5800	-	-
1920 x 1080	4500	2560 x 1440	3996
1280 x 720	3400	-	-
1280 x 720	2400	1920 x 1080	2324
960 x 540	1600	-	-
960 x 540	900	1280 x 720	1114
960 x 540	600	960 x 540	673
768 x 432	300	640 x 360	328
640 x 360	145	480 x 270	201

TABLE 2: AN EXAMPLE OF CONTENT ADAPTED OTT PROFILES COMPARED TO HLS RECOMMENDATION FOR HEVC/H.265.

The examples of Table 1 and Table 2 are not unique. Another content adaptive compression tool might generate another set of profiles perfectly consistent, depending on the codec used and external constraints. The goal of this paper is to provide technical input on content adaptive challenges, directions on how a content adaptive compression system can be implemented, and how AI can be helpful. It will illustrate how the above examples have been generated, as well as how future contents and set of constraints could be handled. This paper is primarily focused on file encoding for VOD, but some of the concepts described also apply to live OTT.

In the rest of the paper, numerical data has been generated using the two contents “Polynésie”, in 8k 50fps HDR PQ and “Tour de France”, in 4k 100fps HDR HLG, kindly provided respectively by The Explorers [4] and A.S.O. [5], and illustrated on Figure 1.

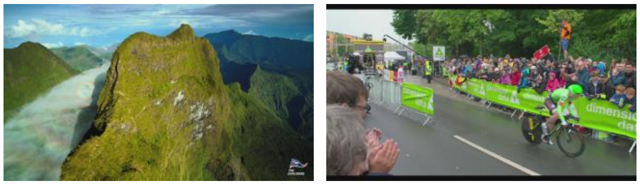


Figure 1: Tests sequences “Polynésie”, in 8k 50fps HDR PQ and “Tour de France”, in 4k 100fps HDR HLG.

General purpose content adaptive framework

I. Quality/rate optimization

The first idea of content adaptive encoding is to provide the best possible quality whatever the bitrate. It is well known that one cannot decrease encoding quality indefinitely. It is necessary to decrease the content resolution at some point. This fact is illustrated on Figure 2. This kind of graph is used classically to illustrate content adaptive encoding [6], [7], though 8k is seldom considered. One can easily notice the bitrate points at which it is relevant to change resolution for optimal quality encoding. Thus, the ideal set of profiles is spread along the maximal convex hull of all the curves. Although very practical, such a graph raises several questions.

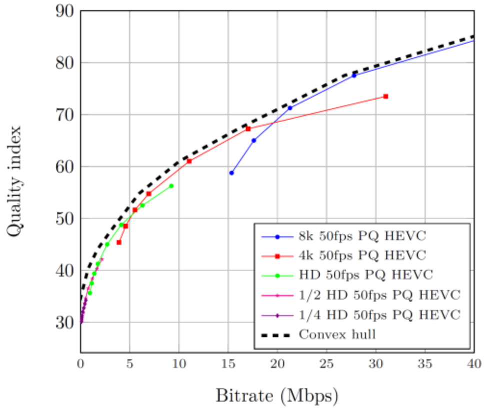
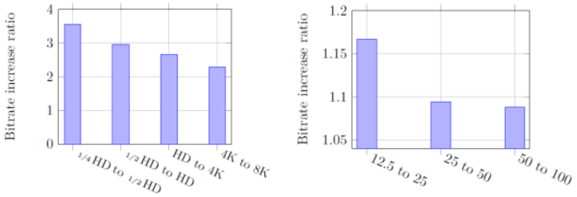


Figure 2: Set of rate-quality curves for resolutions ranging from 480x270 to 8k.

First, the quality index (QI) must be defined, as it is at the very heart of the method. Not only the ability to model actual human perceived quality, but also the universality of the quality index must be questioned. In the case of Figure 2, only resolution is considered because it is the most straightforward example. Perception of resolution is well studied [8] and several quality metrics exhibit consistent behavior against varying resolution, such as scaled-PSNR or VMAF [9]. However, resolution perception is highly dependent on covered visual angle, or in other words, combination of size and distance to the screen. Graphs such as Figure 2 are generally built with fixed to maximum screen size in mind. Hence a lower QI attributed to HD compared to 8k. But if the content is watched on a smartphone, HD might get the highest QI mark, as 4k and 8k would not bring any perceivable improvement. Figure 2 become trickier to draw and use in that case. Nonetheless, the resolution remains the most effective way of acting on bitrate, as illustrated by Figure 3(a). One must note that the bitrate ratios are more prominent on the lowest resolutions. It fits nicely with the fact that resolution increase perception also depends on resolution. The higher the resolution, the less perceivable is a resolution increase. For content adaptive profiles computations, it means that more intermediate resolutions are needed for small than for large resolutions.



(A) **(B)**

Figure 3: HEVC bitrate increase ratio when doubling **(A)** horizontal and vertical resolution for resolutions ranging from 480x270 to 4k, or **(B)** temporal resolution for framerates ranging from 12.5 to 50 fps.

The impact of framerate on perceived quality is far less documented than resolution. Studies exist nonetheless [10]. As for resolution, framerate perception depends on screen size and viewing distance, but also highly on motion characteristics of the content. Generally speaking, very low framerates are easily identified as impairing the perceived quality. Very high framerates, on the other hand, as soon as there is significant motion in the scene, bring eye-catching sharpness and a dramatically increased feeling of reality. This is difficult to transpose into a numerical mark, also consistent with resolution. In the examples of Figure 4, Figure 5 and Figure 6, a quality index has been developed and tuned by ATEME for handling both resolution and framerate. Figure 4 is similar to Figure 2 except for the number of resolutions illustrated, only 3, and the framerate 100fps. Figure 5 is built on the same principle, except that framerate is varying instead of resolution. The result seems consistent. Thresholds appear under which it is better to lower the framerate rather than compressing more. Curves also have a larger overlapping area than for resolution. The impact of framerate on the compression efficiency is actually very different from resolution. Increasing the framerate does not only augment the amount of data, it also makes the frames closer to each other and sharper thanks to shorter shutter speed. Finally, the augmented amount of data is counterbalanced by better temporal prediction during compression. And the higher the framerate, the more prominent this effect, as illustrated by Figure 3(b).

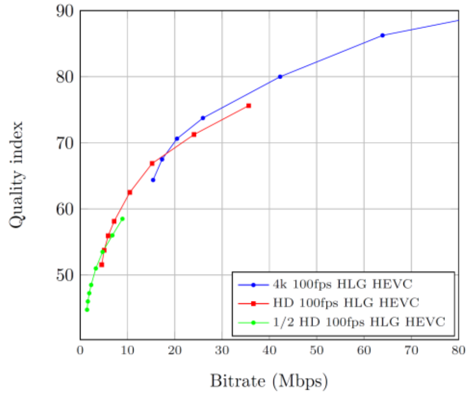


FIGURE 4: SET OF RATE-QUALITY CURVES FOR RESOLUTIONS RANGING FROM 960X540 TO 4K.

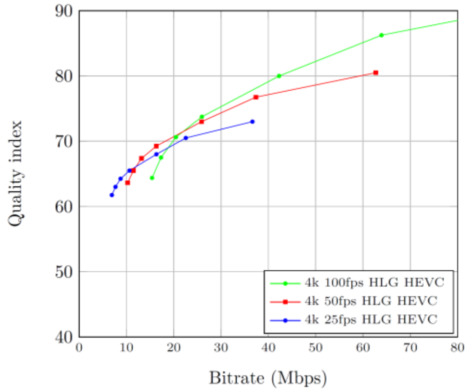


FIGURE 5: SET OF RATE-QUALITY CURVES FOR FRAMERATE RANGING FROM 25 TO 100 FPS.

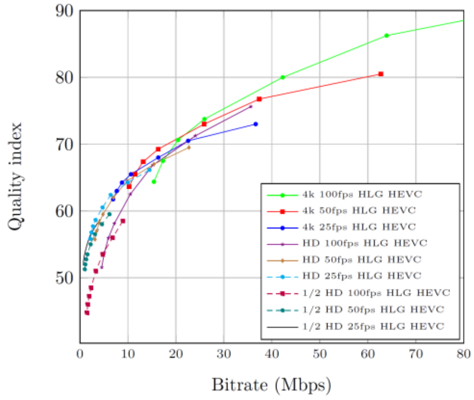


FIGURE 6: SET OF RATE-QUALITY CURVES FOR RESOLUTIONS RANGING FROM 960X540 TO 4K AND FRAMERATE RANGING FROM 25 TO 100 FPS.

Figure 6 illustrates the combination of both resolution and framerate variations. One can spread profiles along the maximal convex hull of the curves and reach a consistent content adapted profiles allocation. However, curves tend to be very close to each other and in some situations, these curves might fall inside each other's error margins. On the other hand, the bitrate gain brought by decreasing framerate is limited, as shown on Figure 3(b).

In summary, framerate effect on perceptual quality doesn't evaluate as compression artifacts or resolution loss, and there is no doubt it would benefit from further studies. Nonetheless high framerate brings valuable perceptual improvement for content with significant motion, at a limited bitrate cost. Therefore, one should avoid decreasing framerate except for small resolutions, on which high framerate brings less perceptual improvement compared to resolution.

All the results presented so far have been computed on HDR contents. But what is the cost of HDR, and is it relevant to consider it in content adaptive profiles optimization? For the cost, Apple's recommendation for HLS [3] considers a bitrate increase of 20%. Our experimental observations are matching this figure overall. Interestingly, it has been observed that this ratio varies depending not only on sequences, but also on bitrate and resolution, so there is room for further analysis here. Regarding perception, viewers should theoretically benefit from HDR whatever the resolution and framerate, as long as the display used is HDR enabled and the ambient lighting is not overly impacting. Nonetheless, it makes sense considering disabling HDR at some point in a profiles ladder in order to convert 20% of bitrate into a slightly higher resolution.

II. Quality model for content adaptive encoding?

Video quality assessment (VQA) is a thoroughly studied topic [11], [12]. Recent advances such as VMAF [9] benefits from the dramatic progress of machine learning, deep learning or generally speaking, artificial intelligence (AI). Even no reference VQA, a particularly challenging task, is making good progress thanks to AI [13]. However, VQA remains an open research area and the focus of a large research community. It is safe to assume there is currently still no perfect universal metric.

It has been shown in the previous section that in the context of content adaptive encoding for OTT application, it is of paramount importance to have, if not a universal metric, at least some reliable ways to evaluate perceived quality of encoding assets in the presence of resolution, framerate and color gamut variations. ATEME's strategy to achieve this goal is to decompose the quality index into several quality descriptors forming together a feature vector. Let's assume for instance that one encodes a 4k / 100fps / HDR asset into an HD / 50 fps / SDR stream at a given rate. Instead of computing a single quality index as commonly considered, a quality vector (QV) is derived containing:

- Encoding quality (EQ), the quality of the encoded stream compared to the actual HD / 50 fps / SDR encoder input.
- Spatial index (SI), the perceived quality loss incurred by resolution down-sampling, as a

function of viewing parameters and content spatial features.

- Temporal index (TI), the perceived quality loss incurred by framerate reduction, as a function of viewing parameters and content motion analysis.
- Dynamic range index (DRI), the perceived quality loss of tone-mapping to SDR as a function of content dynamic range characteristics.
- Color gamut index (CGI), the perceived quality loss of reducing the color gamut as a function of content color characteristics.

In short: QV = < EQ, SI, TI, DRI, CGI>.

Thanks to this problem decomposition, each feature of QV can be studied separately. The full derivation of each index is beyond the scope of this paper, but the principle is the same for each. A training base is defined and annotated manually. Relevant features are extracted from the training base contents, as for instance motion fields for deriving temporal index. A machine learning algorithm is then trained on this data. The problem decomposition and features selection help keeping the learning reasonably deep. It is still possible to train a single quality metric, as a function of QV. The ATEME Quality Index (AQI) is defined as a function of QV and visual angle coverage.

As shown in the previous section, an ideal set of profiles may be chosen from a large set of quality / rate points, lying on an optimal convex hull. This raises two questions.

First, how to generate efficiently a large amount of quality / rate points? Some state-of-the-art approaches actually encode several times the asset to observe a posteriori quality / rate points. This strategy is of course highly CPU intensive. We prefer estimating quality / rate points, thanks to deep learning. Thus, using the previously mentioned learning base, and associated computed QV, it is possible to train a deep learning algorithm to estimate the QV for any input content.

Secondly, how to plot 2D-points and extract some convex hull when the quality has become multi-dimensional? It is indeed much more difficult, unless one forgets about QV and relies on AQI. Instead of that ATEME proposes a different and more general approach, not relying on machine learning this time, but on a more classical trellis optimization.

III. Trellis optimization

Let us first provide a reduced example, without loss of generality. Suppose a trellis has been built, as shown on Figure 7. Each node is labelled with a resolution, a bitrate and an AQI score, respectively. Each vertex represents a valid transition between two profiles. Given any design criteria, one can derive the optimal path in the trellis relative to that criteria. For instance, the upper part highlighted path of Figure 7 maximizes the profiles perceptual quality, while the lower part highlighted path minimizes the overall profiles storage cost. The point is that given a trellis corresponding to an asset, one can generate an optimal set of profiles following any arbitrary constraint. Let's assume that the user requires an 1280x720 profile, then the path corresponding to the minimal storage cost is easily modified accordingly.

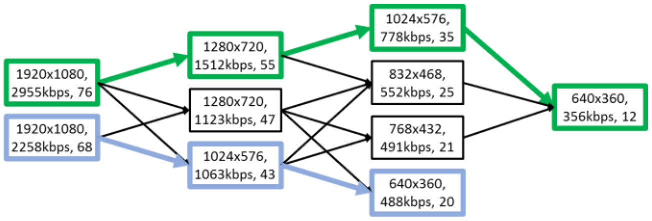


FIGURE 7: CONTENT ADAPTIVE TRELLIS OPTIMIZATION EXAMPLE.

The heart of this optimization lay in the trellis design. The trellis is built such that all paths respect a set of constraints forcing the corresponding set of profiles to be consistent. Basically, one can think about a fully connected trellis built with all the available points repeated on several layers up to a given maximum of layers. A pruning process is then applied to reduce the trellis to only the so-called valid paths. Once again, it is beyond the scope of the paper to describe extensively the whole process. Still, the design criterion allowing to determine the valid nodes and vertices summarize as follows:

- Monotonicity in all dimensions (rate, resolutions, framerate, dynamic range, AQI, QV elements).
- Significant enough rate step, in order to ensure enough margin for network adaptation.
- Seamless quality transition i.e. making sure that when a player transitions from one profile to another, it will not be noticeable.
- Sufficient encoding quality. Basically, any node with an EQ lower than a given threshold will be removed.

IV. Summary

The proposed process is summarized by Figure 8. First, the asset is analyzed thanks to AI. Let us remind here that two AI-based processes are involved. One for a-posteriori with reference quality evaluation and one another for a-priori quality estimation. As a result, a large set of possible rate / quality points is available. Each point stores a rate, a codec, a resolution, a framerate, a dynamic range, an AQI and a QV. Second, a trellis modeling all the possible profiles set is built. An optimal path in the trellis is derived depending on any arbitrary constraint and optimization criterion. Finally, encodings are performed for the selected profiles. One must note that his system handles either CBR or VBR.

This proposed strategy has many advantages. It is reasonably complex, as the initial machine learning problem has been broken down into several smaller problems and trellis model is very efficient. Classical convex hull approach is in fact a sub-product of the model (AQI maximization). The consistency of the set of profiles is guaranteed by construction. And finally, it is highly flexible, as any arbitrary constraint can be integrated, such as mandatory or forbidden profiles, player constraints and so on. All these features make the system highly future-proof.

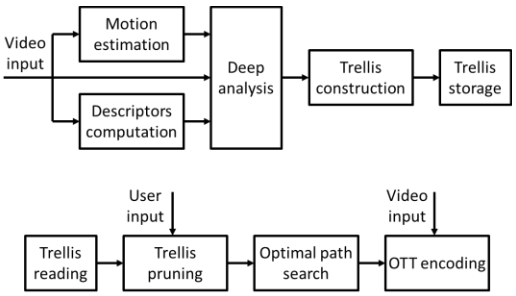


FIGURE 8: ATEME CONTENT ADAPTIVE ENCODING PROCESS.

As an illustration of its flexibility, the framework has been applied on the two sequences used all along this paper. The codec has been fixed to HEVC, and a maximum mandatory resolution and a minimum mandatory bitrate have been provided as constraints. The resolution, framerate and dynamic range have been left to the framework to be tuned automatically. Table 3 presents the result obtained on the sequence Polynésie. The profiles ladder looks usual, except for the resolution spanning a very wide range. Resolution is decreasing smoothly along with bitrate, while framerate is decreased only for the lowest profiles.

Codec	Resolution	Fps	Dynamic range	Bitrate (kbps)
HEVC	7680x4320	50	HDR	17606
HEVC	3840x2160	50	HDR	6924
HEVC	2560x1440	50	HDR	3095
HEVC	1920x1080	50	HDR	1755
HEVC	1280x720	50	HDR	1054
HEVC	960x540	50	HDR	642
HEVC	640x360	25	HDR	383
HEVC	480x270	25	HDR	224

TABLE 3: CONTENT ADAPTED SET OF PROFILES FOR SEQUENCE POLYNÉSIE.

A less usual result is obtained on the Tour de France sequence, as shown in Table 4. This is a sport sequence in which motion plays an important role, combined with a high level of spatial details. On the other hand, the rainy weather brings less emphasis on HDR. These features are identified by the deep learning-based analysis phase. Thus, the automatic profile recommendation tends to favor high framerate, followed by resolution, while the HDR feature is dropped very early as bitrate decreases.

Codec	Resolution	Fps	Dynamic range	Bitrate (kbps)
HEVC	3840 x 2160	100 fps	HDR	20477
HEVC	2560 x 1440	100 fps	HDR	9697
HEVC	1920 x 1080	100 fps	SDR	5859
HEVC	1280 x 720	100 fps	SDR	3649
HEVC	1280 x 720	50 fps	SDR	2381
HEVC	960 x 540	50 fps	SDR	1564
HEVC	960 x 540	50 fps	SDR	1042

TABLE 4: CONTENT ADAPTED SET OF PROFILES FOR SEQUENCE TOUR DE FRANCE.

Additionally, to rate, resolution, framerate and dynamic range adaptation, the proposed framework could even handle codec switching if necessary. It would be also possible to derive a single set of profiles addressing several possible screen sizes.

CONCLUSION

In this paper, an analysis of the general perceptual features of current and up-coming video contents is provided. From that analysis, a generic content adaptive encoding framework is derived, relying both on artificial intelligence and trellis optimization algorithms. The main advantages of the proposed framework are its optimal adaptation to any content, including UHD 8k, HFR and HDR using any codec such as HEVC, AV1 or VVC, and its ability to enforce any arbitrary constraint. As such, the proposed framework is highly future-proof. Future works to be presented includes live content adaptation and player side optimizations.

REFERENCES

[1] "Ultra HD Forum Draft: Ultra HD Forum Phase B Guidelines," https://ultrahdforum.org/resources/phaseb-guidelines-description/, April 07, 2018

[2] "Cisco Visual Networking Index: Forecast and Trends, 2017–2022," https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html, November 26, 2018

[3] "HLS Authoring Specification for Apple Devices," https://developer.apple.com/documentation/http_live_streaming/hls_authoring_specification_for_apple_devices, November 09, 2018

[4] https://theexplorers.com/

[5] https://www.aso.fr/

[6] "Per-Title Encode Optimization," Netflix tech blog, https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2, December 14, 2015

[7] "LightFlow," Epic Labs, https://lightflow.media/, 2018

[8] Zhao, Yin et al. "Video quality assessment based on measuring perceptual noise from spatial and temporal perspectives." IEEE Transactions on Circuits and Systems for Video Technology 2011 : 1890-1902.

[9] "VMAF: The Journey Continues," Netflix tech blog, https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12, October 25, 2018

[10] R. M. Nasiri, J. Wang, A. Rehman, S. Wang and Z. Wang, "Perceptual quality assessment of high frame rate video," 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP), Xiamen, 2015, pp. 1-6

[11] Chikkerur, Shyamprasad et al. "Objective video quality assessment methods: A classification, review, and performance comparison." IEEE Transactions on Broadcasting 57.2 PART 1 (2011) : 165-182

[12] Vranješ, Mario, Snježana Rimac-Drlje, and Krešimir Grgić. "Review of objective video quality metrics and performance comparison using different databases." Signal Processing: Image Communication 28.1 (2013) : 1-19

[13] C. Wang, L. Su and W. Zhang, "COME for No-Reference Video Quality Assessment," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 232-237