MPEG-5 part 2: Low Complexity Enhancement Video Coding (LCEVC): Overview and performance evaluation

Conference Paper · August 2020

DOI: 10.1117/12.2569246

MPEG-5 part 2: Low Complexity Enhancement Video Coding (LCEVC): Overview and performance evaluation

Guido Meardi[a], Simone Ferrara[a], Lorenzo Ciccarelli[a], Guendalina Cobianchi[a], Stergios Poularakis[a], Florian Maurer[b], Stefano Battista[c], Ahmad Byagowi[d]

[a]V-Nova, London, United Kingdom; [b]RWTH Aachen University, Aachen, Germany; [c]Università Politecnica delle Marche, Ancona, Italy; [d]University of Manitoba, Winnipeg, Canada

ABSTRACT

Low Complexity Enhancement Video Coding (LCEVC) is a new MPEG video codec, currently undergoing standardization as MPEG-5 Part 2. Rather than being another video codec, LCEVC enhances any other codec (e.g. AVC, VP9, HEVC, AV1, EVC or VVC) to produce a reduced computational load and a compression efficiency higher than what is achievable by the enhanced codec used alone for a given resolution, especially at video delivery relevant bitrates. The core idea is to use a conventional video codec as a base codec at a lower resolution and reconstruct a full resolution video by combining the decoded low-resolution video with up to two enhancement sub-layers of residuals encoded with specialized low-complexity coding tools.

Keywords: Low Complexity Enhancement Video Coding (LCEVC), MPEG-5 Part 2, video coding, standard, enhancement, software codec, video compression, multi-resolution video coding.

1. INTRODUCTION

LCEVC responds to the need for higher quality video in bandwidth-constrained scenarios, with the sustainability of processing complexity as a target. LCEVC design allows purely software implementation for low power hardware found on a broad range of standard devices, side-stepping the requirement for dedicated hardware blocks in the silicon. This software implementation option provides a backward-compatible capability to upgrade video services based on minimum legacy hardware, with old and new codecs alike.

Since LCEVC is codec agnostic, it can simplify the transition from an older generation codec to a newer one during upgrades, whilst remaining beneficial even after completion of the upgrade. The codec-agnostic ability provides the LCEVC user with the capacity of rapid improvement in the quality of the offered service to all users, and at the same time maintains a degree of performance improvement also as the service starts migrating towards next-generation codecs.

LCEVC uses a multi-layer approach and scaling tool. Therefore, on the traditional attributes of scalability, LCEVC it is not considered as a scalable codec; in particular, it does not satisfy the requirements of codecs such as SVC or SHVC, and vice versa.

Applications such as over-the-top (OTT) streaming, live video feeds from mobile devices, videoconferencing over public internet are always battling with the issue of bandwidth. Prior to LCEVC, the only option for dealing with bandwidth constraints was to either switch to a codec with a higher compression rate or reduce the video resolution of the stream.

Adopting a new codec can indeed produce material improvements in compression efficiency but requires the end-user devices to support that codec. For high-resolution videos and battery-powered devices, that means hardware support in the silicon. With the current improvement rates of consumer electronics, from the moment a silicon manufacturer decides to implement a codec in hardware, approximately a decade is needed until large majority of video-capable devices can adopt that codec. For example, eight years after completion of the standardization of HEVC, a relevant portion of the installed base of video-capable devices – including devices sold in 2020 – does not support HEVC. Of course, newer codecs can be used much earlier for portions of a service, serving new devices that support the codec of choice, but large portions of an operator's cost base and revenue base are still going to be predicated on older hardware-based codecs for a long period.

Another option to cope with bandwidth constraints is to downsample a video and encode it at a lower resolution. It is known in the art [1][2][3] that below a certain quality threshold (which depends on every specific self-similar video sequence, or "shot"), a lower resolution encode can produce a higher overall video quality than that of a full resolution encode, trading off better picture consistency with loss of detail. The shell that – for a metric of choice – outlines the highest boundary of all the RD-curves of a shot (i.e., a short self-similar video sequence) at various different resolutions is commonly called "convex hull" for that shot-metric combination and defines for each given bitrate the resolution that produces the highest quality.

Downsampling a video and encoding at low resolution is a commonly adopted strategy for OTT streaming as well as videoconferencing. It allows to reduce the risk of "quality collapse", with the price of loss in resolution, loss of details and subsampling impairments. These compromises may reduce the perception of depth or ability to identify elements in the video, with ultimate impact on Quality of Experience, user engagement and ultimately the returned revenue. It should be noted that, when adopting this strategy, it is necessary to downsample the entire video, including areas of the picture (e.g., tickers, small-print text, HUDs, logos, advertising, facial features, a ball) that may be relevant to creative intent and quality of experience.

LCEVC was designed to provide a third option – "orthogonal" and compatible with the previous two, rather than alternative – in order to cope with bandwidth constraints while maintaining visual quality (or vice versa upgrade visual quality without substantially increasing bandwidth requirements): enhancing the codec of choice with LCEVC. Low computational complexity played a fundamental role in the design of the codec, since the whole point of LCEVC lies in its possibility to efficiently leverage spare general-purpose processing power with sustainable power consumption, reducing the overall processing complexity of the combination of base codec at lower resolution plus LCEVC.

This manuscript illustrates the key coding tools and features of LCEVC, provides performance results in combination with a range of codecs and explores some of the reasons why LCEVC works the way it does, including an analysis of where it works best and where it works worst.

2. OVERVIEW OF LCEVC

2.1 Bitstream structure

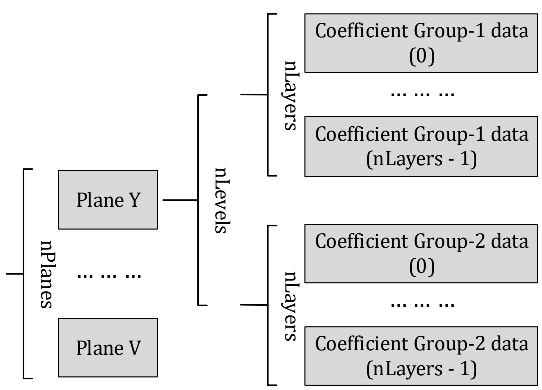


Figure 1. Example scheme of an LCEVC bitstream

An LCEVC encode entails two separate bitstreams, namely a base bitstream produced by a base encoder conformant to its associated specification (e.g. H.264/AVC, H.265/HEVC, H.266/VVC) and an enhancement bitstream (the LCEVC bitstream) produced by the enhancement encoder and conformant to specification ISO/IEC 23094-2 (which at the time of this paper is in draft international standard (DIS) status).

The LCEVC bitstream is structured in a specific order, depending on the chosen configuration parameters. A simplified structure is visualized in Figure 1. The LCEVC encoder can be set to enhance all of the three available planes (luma and chroma) or the luma plane, only. Within each encoded plane, up to two levels can be present. The first level (enhancement sub-layer 1) is used to encode transformed and quantized residuals before applying the final upscaling. The second level (enhancement sub-layer 2) is then added after the final upscaling, meaning at the same resolution as the overall output sequence. Each of these two sub-layers is split into coefficient groups. Depending on the chosen transform type, with a kernel size of 2×2 or 4×4 samples, either 4 or 16 groups are present within a level. Each coefficient group contains the corresponding transform coefficients.

Furthermore, temporal signaling is added as an additional group to enhancement sub-layer 2 if the temporal prediction is active. This group contains information whether residuals from the previous frame, stored in a temporal buffer, are used for prediction on a block-by-block basis.

2.2 Encoder

The structure of an LCEVC encoder is depicted in Figure 2. The encoding process can be divided into three main steps.

Firstly, the input sequence is downscaled using a non-normative downscaler. Depending on the chosen configuration, the downscaling can be applied up to two consecutive times. The video, now at a lower resolution than the input sequence, is fed into the base encoder (e.g. H.264/AVC, H.265/HEVC, H.266/VVC). This process is not further specified in LCEVC: any encoder that produces a decodable bitstream can be used.

Using a normative upscaler, which allows the use of different and content-adaptive kernels as well as the optional encoder-signalled activation of a non-linear corrector called Predicted Residuals, the upscaled base reconstruction then functions as an input for the second step of the LCEVC encoding process. The enhancement sub-layer 1 (L-1) residuals are created by subtracting the downscaled input sequence and the base reconstruction. These residuals, which are typically sparse (e.g. edges or details), are transformed, quantized and encoded resulting in coefficient groups as discussed in subsection 2.1. The transform used in LCEVC has a simple structure and uses a small kernel of size 2×2 or 4×4. This allows to both efficiently code sparse information and parallelize the transforms, since individual blocks are not dependent on other blocks within a frame. A linear quantizer, which may include an adaptive dead-zone, is used to further process the transform coefficients. The entropy encoder, which consists of a run-length encoder (RLE) and an optional prefix encoder (Huffman encoder), processes the quantized transform coefficients and creates the coefficient groups for sub-layer 1.

The inverse tools of the quantization and transform are applied, with adaptive and optionally encoded-signaled asymmetric dequantization, creating the sub-layer 1 reconstruction. Additionally, an L-1 filter can be added which functions as a simple deblocking filter.

Finally, the sub-layer 1 reconstruction is upscaled to full resolution and subtracted from the original input sequence. The resulting enhancement sub-layer 2 (L-2) residuals are fed into the temporal prediction. LCVEC uses a zero-motion vector temporal scheme which operates on a block-by-block basis. The residuals from the previous frame are stored in a temporal buffer and are added to the L-2 residuals in case the temporal prediction is activated. To reduce the signaling overhead in e.g. a fast-moving sequence, where this zero-motion vector scheme would likely not be beneficial, the temporal prediction can be disabled for a group of pixels of size 32×32, as well as for an entire frame with a single bit. The temporal signaling, containing the information whether the temporal prediction is active for a specific transform block, is entropy encoded and included as a temporal layer in the LCEVC bitstream. The residuals are transformed, quantized and encoded using the same tools as explained for sub-layer 1. The quantizer can use different quantization parameters for the two enhancement sub-layers allowing to balance the impact of the two sub-layers and to decide where to add more details.

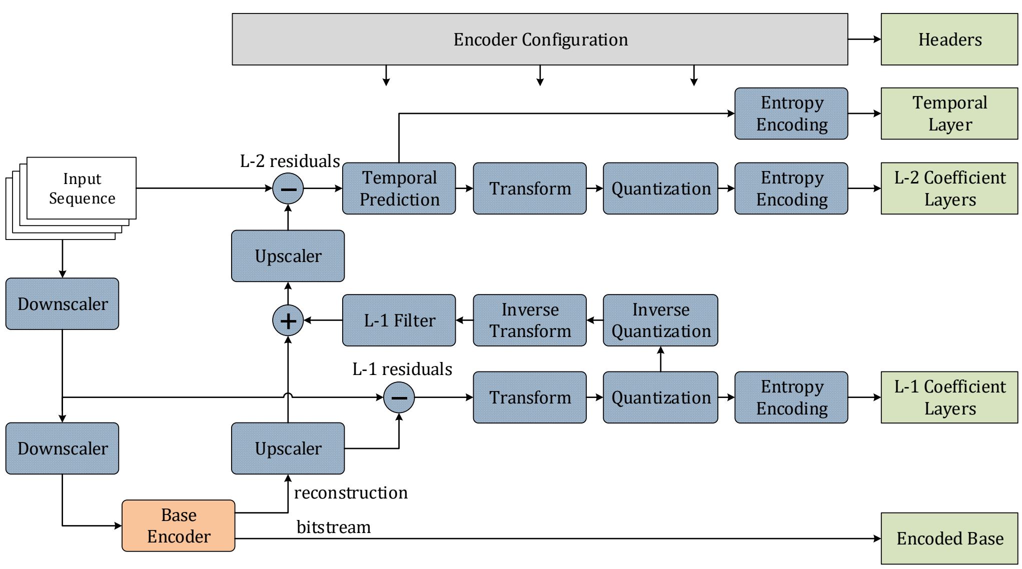


Figure 2. Structure of an LCEVC encoder

2.3 Decoder

The functionality of a normative LCEVC decoder is standardized in [4], and its structure is visualized in Figure 3. Similar to the encoder, explained in subsection 2.2, three main steps are visible: the base decoding and the corrections in enhancement sub-layers 1 and 2.

The decoding process is highly parallelizable, and amenable to both SIMD- and GPU-accelerated processing. The base video decoder is independent from the LCEVC enhancement part. Additionally, the two enhancement sub-layers can be reconstructed in parallel as well. No inter-block dependencies are present within a frame making the inverse transforms independent from other blocks. If the temporal prediction is active, the enhancement sub-layer 2 reconstruction is dependent on the temporal buffer which stores residuals from the previous frame.

Using the same (or inverse) tools explained in subsection 2.2, the reconstructions from the three main decoding steps are achieved. Those are combined using two normative upscalers and additions resulting in the decoded output sequence.

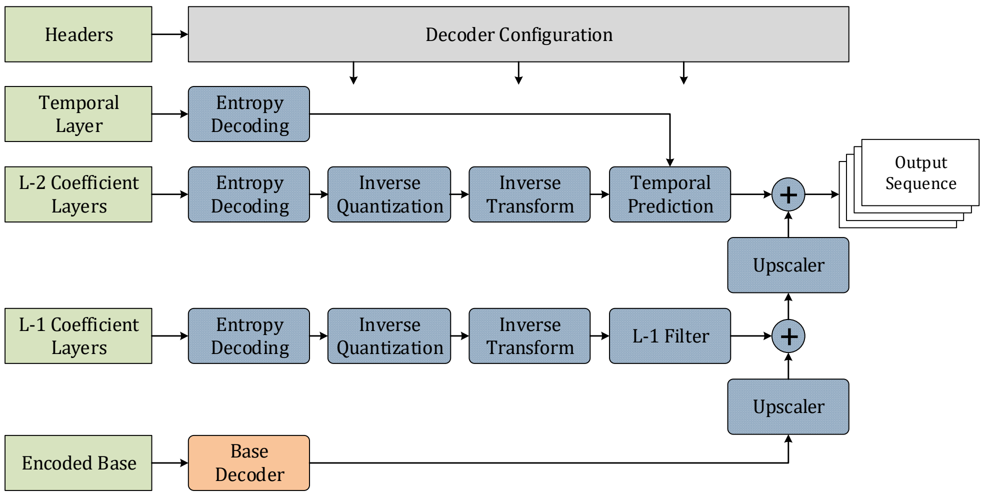


Figure 3. Structure of an LCEVC decoder

2.4 Flexibility and extensibility

LCEVC is not a standalone codec but requires a base codec to operate. The base codec does not need to fulfill any requirements in order to be used in conjunction with LCEVC. This allows the use of any codec as a base codec, current and future ones, making it possible for LCEVC to be used with any existing and future codec without the need to revisit the LCEVC standardization. In addition, LCEVC offers the flexibility to customize most of the decoding tools.

Firstly, the dequantization uses a quantization matrix which contains the actual quantization step widths to be used to decode each coefficient group. For the calculation of this matrix, default coefficients are preset in LCEVC. In addition, custom coefficients can be signaled in the bitstream creating the opportunity to modify the quantization process on a frame-by-frame basis to better work for specific sequences. The custom coefficients can either be signaled for only one or both enhancement sub-layers.

Secondly, there are four preset upscalers which can be selected by the encoder, ranging from a simple upscaling process (nearest) to a more complex one (modified cubic interpolation). In addition, a fifth upscaler can be used with customizable kernel coefficients. This allows to adapt the upscaling process to improve the quality of the upscaled enhancement sub-layer 1 reconstruction for sequences where the adaptive cubic interpolation is beneficial.

Third, LCEVC incorporates the ability to signal to the decoder post-processing operations such as statistical dithering, which can be useful to reduce banding/aliasing impairments and efficiently reconstruct film/camera grain. The in-loop and out-of-loop functionality of LCEVC can be further extended with the insertion of local user data: LCEVC allows to efficiently signal a specified number of bits at transform block level. The bits are embedded in one of the coefficient groups of enhancement sub-layer 1. User data bits, ignored by a reference decoder, can be optionally processed by a media player to control additional parameters in a system, allowing for novel backward-compatible enhancement extensions.

3. PERFORMANCE ANALYSIS USING MPEG TEST MODELS

3.1 Test setup

In order to show that LCEVC enables the targeted performance improvements over a given base codec, we ran several tests on LCEVC using H.264/AVC, H.265/HEVC and H.266/VVC as video coding standards for the base codec.

In particular, the tests were performed using the LCEVC reference implementation in its current version LTM 4.1 [5]. In its current implementation, LTM 4.1 uses as a base codec the reference software for the respective standard: either JM 19.0 for H.264/AVC [6] or HM 16.18 for H.265/HEVC [7].

In order to test LCEVC also over H.266/VVC, the same approach described above for H.264/AVC and H.265/HEVC was adopted. The input sequence downsampled with the Lanczos3 filter included in LTM 4.1 was encoded using the reference software VTM 9.0 for H.266/VVC [8]. The base encoded sequence was then enhanced using LTM 4.1.

For all the above reference implementations, we used their corresponding Random Access configurations.

3.2 Compression efficiency results

The performance results are shown using the two objective metrics Peak Signal-to-Noise Ratio (PSNR) and Video Multimethod Assessment Fusion (VMAF) [9][10]. The latter is operated using the default model (v0.6.1) for HD sequences and the 4K model (4k_v0.6.1) for UHD sequences.

All the results compare a full resolution video encoded using the anchor codec (H.264/AVC or H.265/HEVC) against a full resolution video encoded using LCEVC. When comparing against an AVC anchor, LCEVC would use AVC as the base codec. When comparing against an HEVC anchor, LCEVC would use HEVC as the base codec.

The video data sets listed in Table 1 has been used.

Table 1. Test Sequences

Test Sequence	Sequence Name	Resolution	Frame Rate	Base Codec Tested
A1	Campfire	3840x2160	30	AVC, HEVC
A2	ParkRunning3	3840x2160	50	AVC, HEVC, VVC
A3	FoodMarket4	3840x2160	60	AVC, HEVC
A4	Fortnite (Part 1)	3840x2160	60	AVC, HEVC
A5	Cat Robot	3840x2160	60	AVC, HEVC, VVC
A6	Daylight Road	3840x2160	60	HEVC, VVC
A7	Flying Birds HDR	3840x2160	60	HEVC, VVC
A8	Sunset Beach HDR	3840x2160	60	HEVC, VVC

In a first test, LCEVC was tested using AVC and HEVC as base codecs and compared with the corresponding full resolution anchors. The same quantization parameters (QP) were used for each sequence. In particular, for the anchors the following QPs were used: 42, 37, 32 and 27. For the base encodes, the following QPs were used: 34, 30, 25 and 21. Both enhancement sub-layer 1 and sub-layer 2 were used for the LCEVC bitstream. For this test, also formal mean opinion scores (MOS) (average and CI) were produced, according to ITU-R Recommendation BT.500 [11].

Tables 2 and 3 provide the average coding performances of LCEVC using the Bjøntegaard metric (BD-rate) [12] using AVC and HEVC as a base codec, respectively.

Table 2. BD-rates for Test 1 (LCEVC over AVC vs. AVC)

Test Sequence	Base Codec	BD-rate PSNR	BD-rate VMAF	BD-rate MOS
Campfire	AVC (JM)	−21.90%	−29.30%	−33.57%
Cat Robot	AVC (JM)	−24.44%	−37.56%	−37.80%
FoodMarket4	AVC (JM)	−31.45%	−41.05%	−48.41%
Fortnite (Part 1)	AVC (JM)	−20.52%	−36.86%	−51.61%
Park Running 3	AVC (JM)	−39.49%	−50.01%	−59.60%
Average		***−27.56%***	***−38.95%***	***−46.20%***

Table 3. BD-rates for Test 1 (LCEVC over HEVC vs. HEVC)

Test Sequence	Base Codec	BD-rate PSNR	BD-rate VMAF	BD-rate MOS
Campfire	HEVC (HM)	−19.99%	−31.42%	−41.17%
Cat Robot	HEVC (HM)	6.40%	−14.20%	−29.83%
FoodMarket4	HEVC (HM)	−0.65%	−18.27%	−24.32%
Fortnite (Part 1)	HEVC (HM)	22.45%	−7.92%	−19.15%
Park Running 3	HEVC (HM)	−20.04%	−34.29%	−43.90%
Average		***−2.37%***	***−21.22%***	***−31.68%***

When compared with JM and HM, LCEVC provides an average VMAF gain of 39% and 21%, respectively. Similar gains are observed when looking at average MOS – namely over 46% over JM and over 31% over HM.

In Figure 4 we show full distortion curves for Park Running (Figure 4a) and Fortnite (Figure 4b), i.e. those sequences for which the BD-rate gain is highest (Park Running) and lowest (Fortnite)

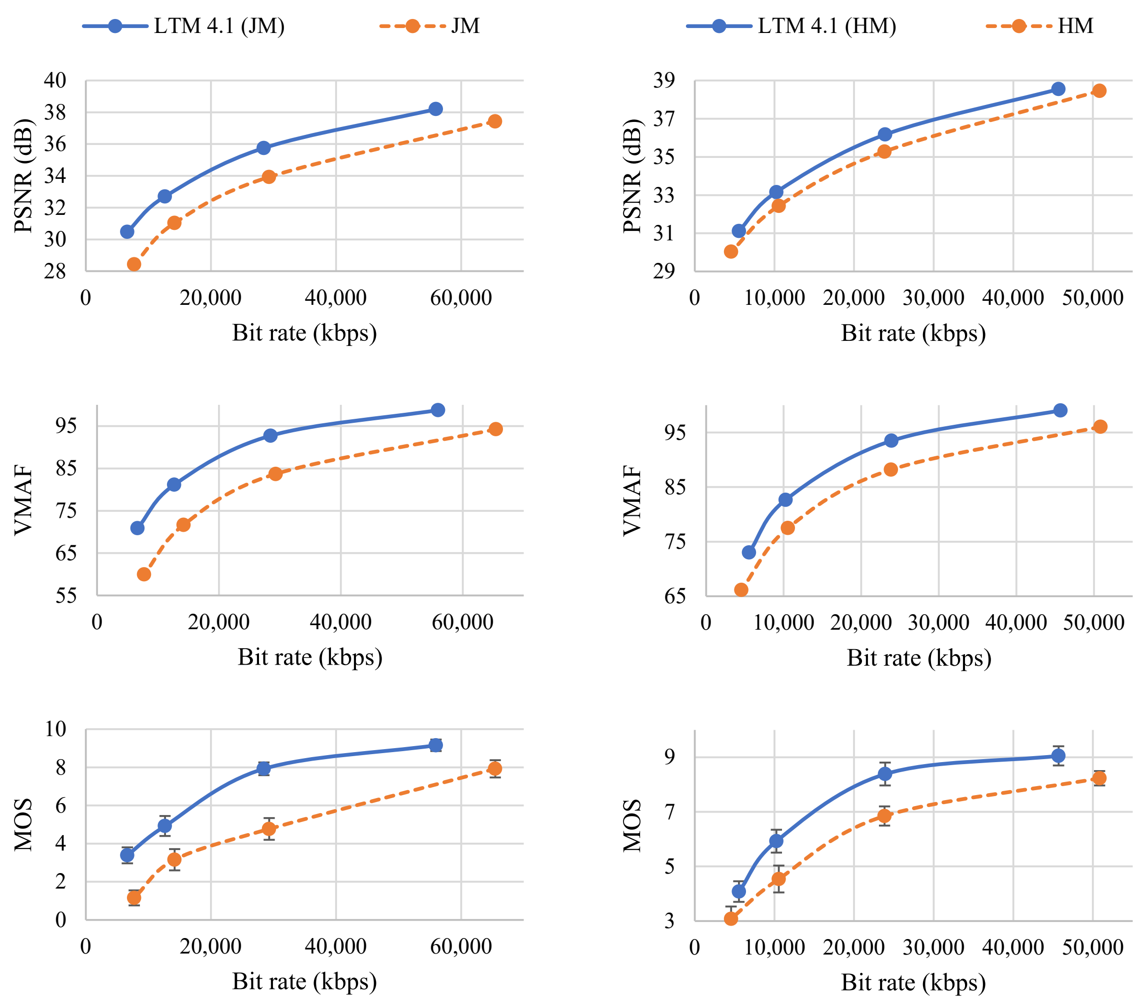


Figure 4a. Distortion curves (PSNR, VMAF and MOS) for Park Running
(left column using AVC as base, right column using HEVC as base)

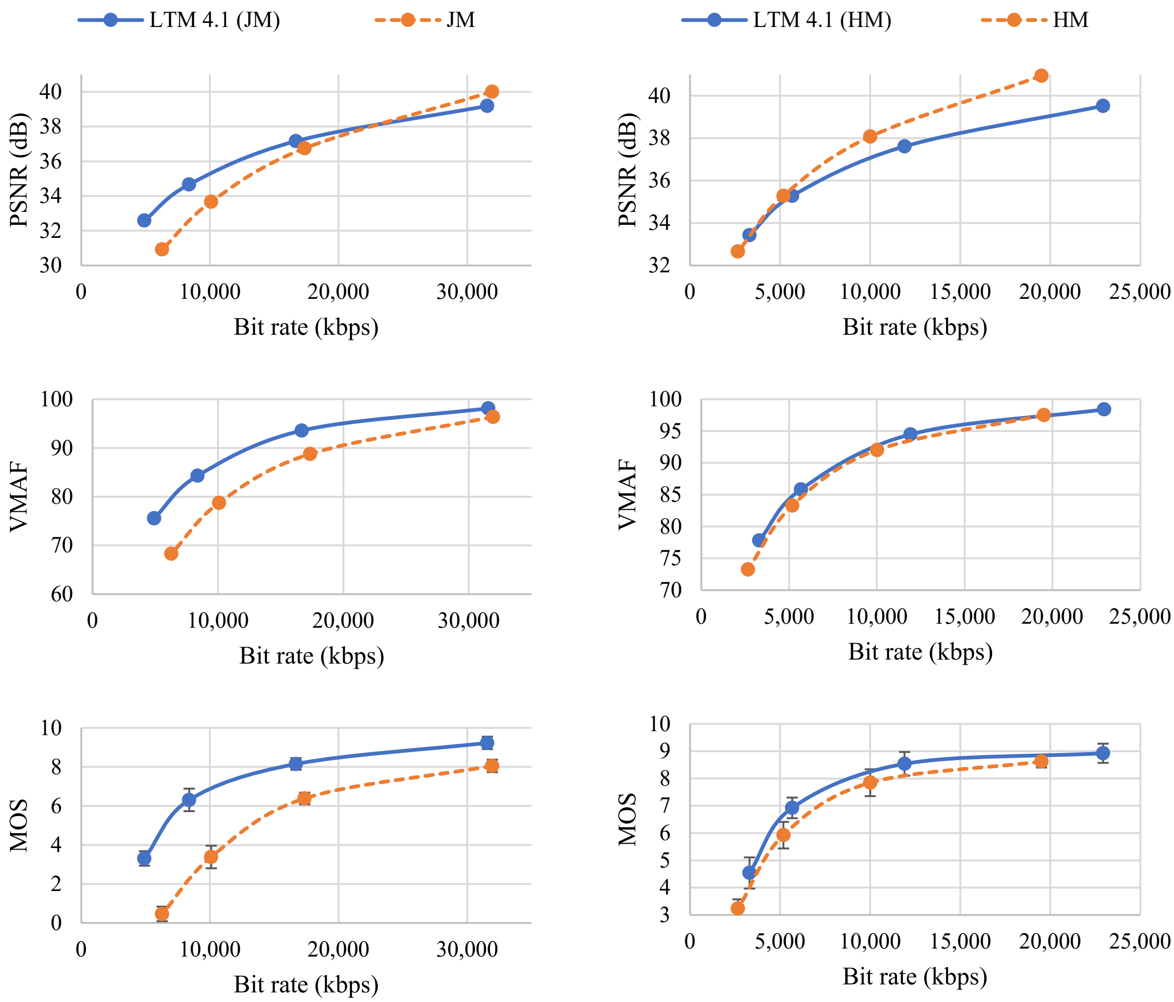


Figure 4b. Distortion curves (PSNR, VMAF and MOS) for Fortnite
(left column using AVC as base, right column using HEVC as base)

In a second test, we tested LCEVC using VVC as base codec and comparing it with the corresponding full resolution anchor. Notice that due to these being among the very first tests of LCEVC use with VVC, no VVC-specific calibration work was performed to suitably set LCEVC for optimal performance with VVC, so we expect potential for improvement. The same quantization parameters were used for each sequence. In particular, for the anchors the following QPs were used: 42, 37, 32, 27 and 22. For the base encodes, the following QPs were used: 34, 30, 26, 22 and 18. Only sub-layer 2 was used for the LCEVC bitstream. To reduce the processing time, all tests have been performed on 65 pictures, then scaling the bitrate consistently with the frame rate (by a factor 50/65 or 60/65). The number of pictures has been chosen as a multiple of 1 Intra period and 4 GOP periods.

Table 4 provides the average coding performances of LCEVC using the Bjøntegaard metric using VVC as a base codec.

Table 4. BD-rates for Test 1 (LCEVC over VVC vs. VVC)

Test Sequence	Base Codec	BD-rate PSNR	BD-rate VMAF
Daylight Road	VVC (VTM 9.0)	35.02%	−3.13%
Cat Robot	VVC (VTM 9.0)	26.15%	−6.63%
Park Running	VVC (VTM 9.0)	6.01%	−6.07%
Flying Birds	VVC (VTM 9.0)	3.22%	−9.88%
Sunset Beach	VVC (VTM 9.0)	33.04%	−8.83%
Average		***20.69%***	***−6.91%***

From Table 4 it can be seen that, when compared with VTM, LCEVC provides an average VMAF gain of almost 7%.

3.3 Analysis of the correlation between subjective and objective metrics

We have further analyzed the data from these tests to determine the correlation between formal subjective MOS scores and objective metrics. This analysis is meant to provide some more insight in the tests, as well as to study how the objective metrics can accurately predict the subjective performances, particularly for LCEVC.

In a first set of analysis, we compared the BD-rates for PSNR and VMAF against the BD-rates for MOS scores.

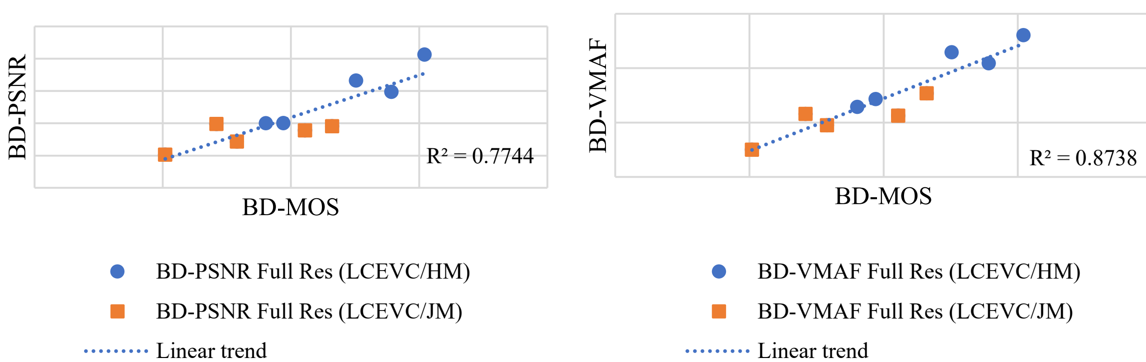
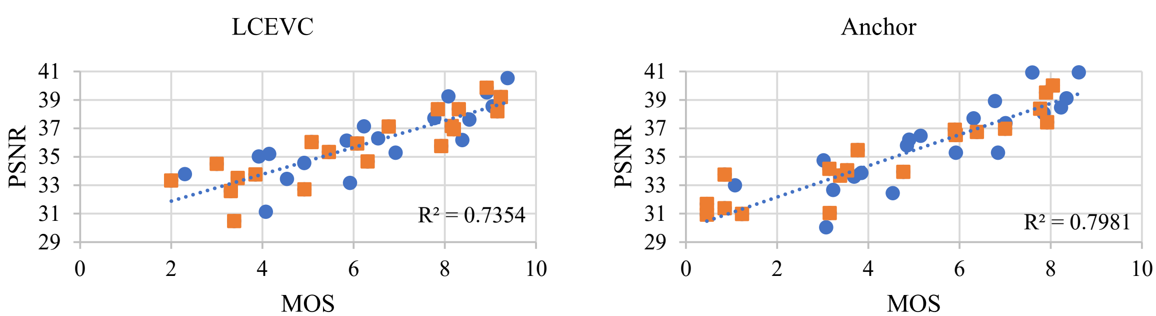


Figure 5. Correlation between BD-PSNR and BD-MOS (left) and BD-VMAF and BD-MOS (right)

In a second set of analysis, we compared the PSNR and VMAF values obtained in the first test against their respective MOS scores. In particular, we compared the values for LCEVC and the values for the anchors separately, to understand how the predictive ability of objective metrics differs across different codecs.



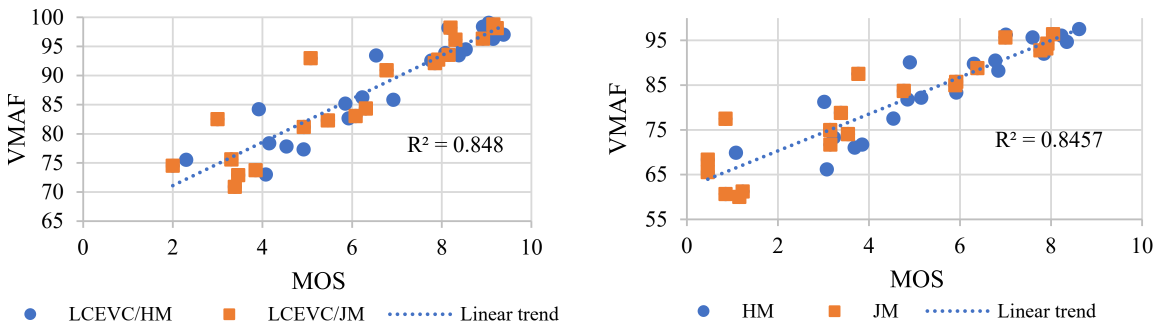


Figure 6. Correlation between PSNR and MOS (top) and VMAF and MOS (bottom)

From Figures 5 and 6 above, the following observations are made:

- for comparison against full resolution anchors, BD-VMAF has a stronger correlation to BD-MOS than BD-PSNR – with a coefficient of determination R^2 of 0.8738; however, despite better overall correlation with MOS also VMAF seems to structurally under-estimate the BD-Rate benefits provided by LCEVC;
- PSNR scores for LCEVC are less correlated to MOS scores than they are for the anchors. This seems to suggest that whereas for the anchors PSNR may correlate reasonably well with MOS, the same may not be true for LCEVC. Especially, the slopes of the correlations are markedly different, with LCEVC on average achieving high MOS scores at ca. 1.5-2 dB lower than the anchors. As a consequence, due to the very different relationship between PSNR and MOS, big divergence with formal MOS scores may arise when we attempt to compare LCEVC and the anchors by means of PSNR comparisons. It should be noticed that the different behavior of LCEVC with PSNR may be partly "inherent" to the codec and partly due to the absence – differently to JM and HM – of any MSE-specific rate-distortion optimizations (RDO) in the LTM: in fact, as shown in section 5, PSNR scores of LCEVC encodes are typically higher at low bitrates, lower at medium bitrates and higher again at very high bitrates, highlighting a different "behavioral pattern" with PSNR;
- in relation to VMAF, VMAF scores have a more similar correlation to the MOS scores for both LCEVC and the anchors (although still with slightly better $R^2$for the anchors than for LCEVC), with more similar slopes.

3.4 Complexity results

A third test has been performed to study the low complexity nature of LCEVC. In particular, using the same setting used for the second test (see 3.2), we measured and compared the encoding times of LCEVC over HEVC, LCEVC over VVC, HEVC and VVC.

In Figure 7 we show a scatter plot of PSNR and VMAF values for every tested point and every tested sequence against the actual encoding times (shown in seconds on a logarithmic scale).

In Figure 8 we show a scatter plot of the BD-PSNR and BD-VMAF for every tested sequence against the encoding times normalized vs the encoding time for HEVC. In particular, the encoding times are the average encoding times for a given sequence and are normalized against the average encoding times for the same given sequence, so that for example a value of 2 means that it is two times the encoding time required to encode the same sequence with HEVC.

As it can be seen, LCEVC significantly reduces the encoding time when compared against the full resolution codec, whilst at the same time improving the compression performance. With AVC the MOS performance is improved by about 35-60%, whilst the encoding time is reduced by ca. 50%. With HEVC the MOS performance is improved by about 20%-40%, whilst the encoding time is reduced by 60%-70%. With VVC – where formal MOS tests have not been performed yet, but gains are expected to remain material - the encoding time is reduced by 80% to between 2x and 3x that required for HEVC, maintaining - and in fact increasing - the gains versus HEVC.

As such, LCEVC confirms satisfaction of its requirements as cross-generational codec-agnostic enhancement, able to provide material compression-complexity gains to past, current and future codecs. Its amenability for power-efficient software processing makes it suitable to be deployed in combination with existing codec via software on legacy devices, whilst possibly also facilitating the adoption of new codecs, via reduction of the dedicated silicon gates/area/power necessary to enc-decode a given resolution.

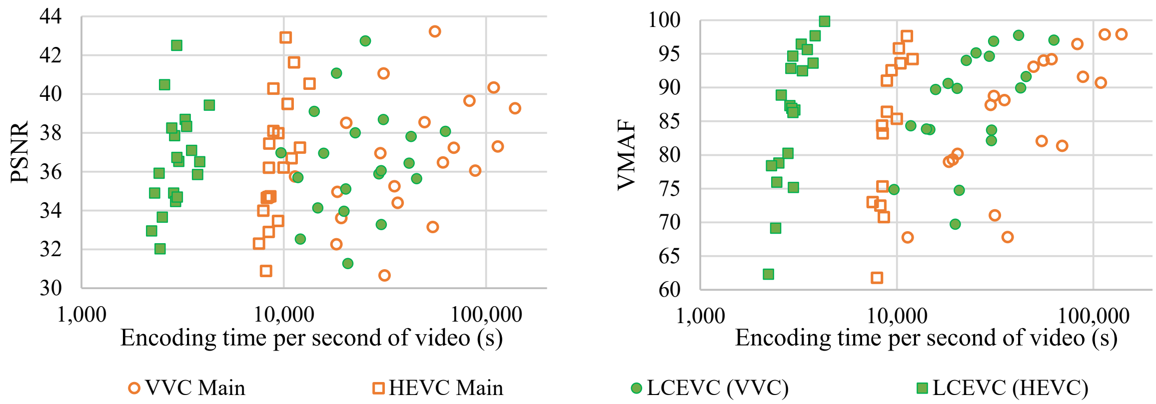


Figure 7. Comparison of PSNR and VMAF values vs. encoding times for LCEVC over HEVC, LCEVC over VVC, HEVC and VVC

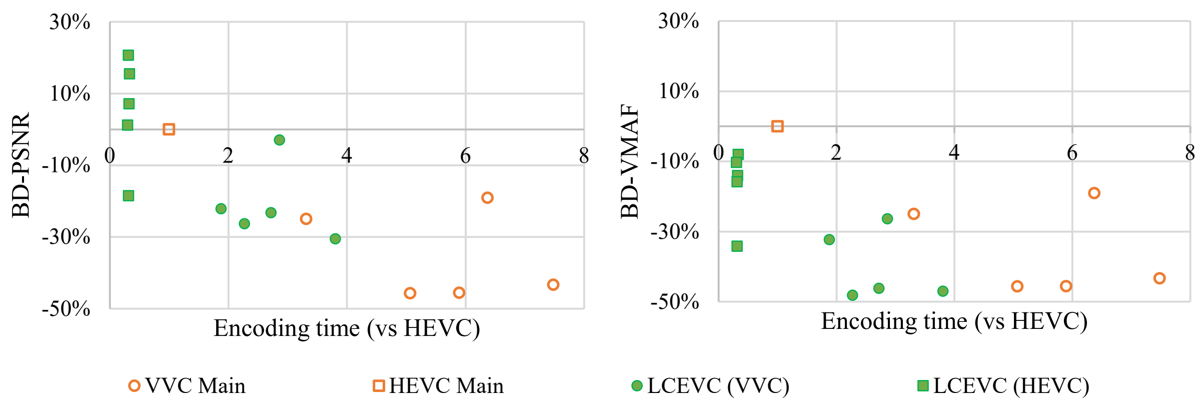


Figure 8. Comparison of BD-PSNR and BD-VMAF values vs. encoding times (normalized over HEVC) for LCEVC over HEVC, LCEVC over VVC, HEVC and VVC

4. PERFORMANCE ANALYSIS USING OPTIMIZED IMPLEMENTATIONS

4.1 Test setup

A further test was also performed with an optimized LCEVC implementation provided by V-Nova, using as both anchor and base codec x264 with preset “slow”. The default constant rate factor (crf) quality setting was used, both at full resolution and at a quarter resolution.

To further validate LCEVC performance with a larger dataset, Table 5 below includes the results achieved when encoding the Netflix El Fuente test set at 1080p (140 short sequences). LCEVC showed BD-rate gains for the vast majority of clips (100% for VMAF, 92% for PSNR), with material average and median gains.

Table 5. Coding performance comparison of LCEVC (optimized implementation) over AVC anchor (x264), variable bit rate

El Fuente HD dataset	PSNR	VMAF
% of sequences with LCEVC BD-rate < 0%	92%	100%
Average LCEVC BD-rate	−27.15%	−47.14%
Median LCEVC BD-rate	−30.72%	−48.06%

Plotting the convex hull of LCEVC-enhanced encodes (1080p, 720p, 540p, 432p, 360p) versus that of native encodes shows that that "convex hull BD-rates" of LCEVC achieve gains that are broadly similar to those obtained by comparing full-resolution encodes. For instance, Figure 9 shows the convex hull comparison for sequence ElFuente_001076_001181.

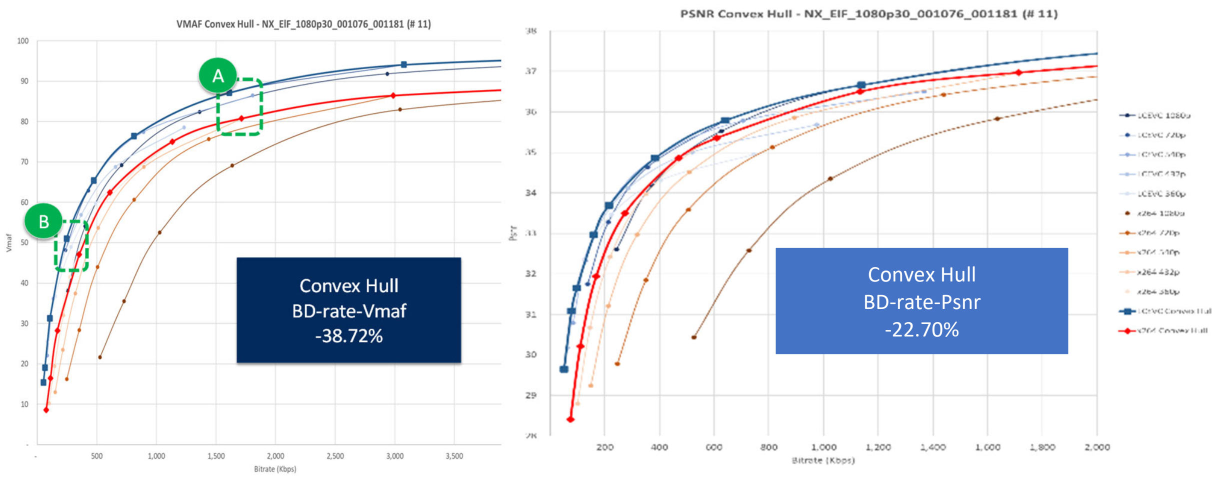


Figure 9. Convex hull comparison for sequence ElFuente_001076_001181.

The quality improvements of LCEVC across the convex hull RD-curve are confirmed by visual inspection, where they manifest as a combination of higher picture consistency (thanks to relatively lower QPs for a same bitrate point) and higher amount of details (thanks to possibility to maintain a relatively higher resolution for a same rate point).

x264: 540p, CRF 23, 1757 kbps **LCEVC x264:** 720p, CRF 27, 1699 kbps

Figure 10. Examples of operating points in the convex hull where LCEVC enables higher visual quality and resolution at lower overall bitrate

Given the broad amount of sequences in the El Fuente test set, we ran statistical analysis to identify patterns for where LCEVC works best and where it works worst. First of all, we noticed no statistically significant correlation between BD-rates and temporal complexity of the sequence as calculated per ITU P.910 recommendations:

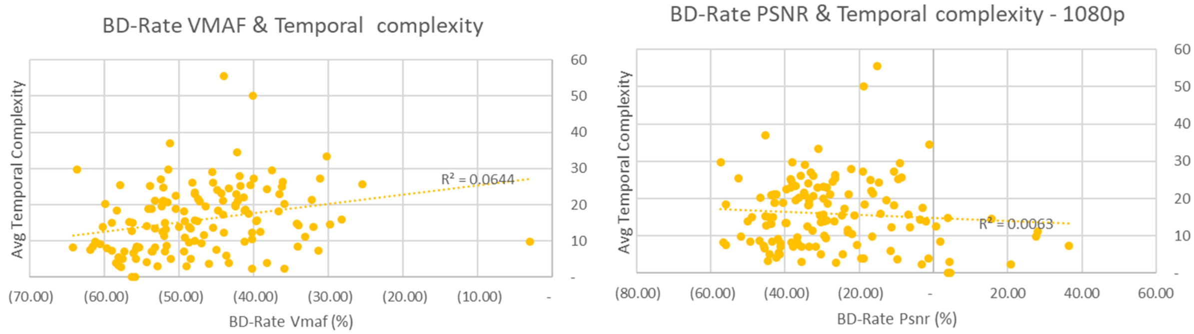


Figure 11. Correlation (or lack thereof) between temporal complexity and BD-Rates.

We found instead some positive correlation with spatial complexity:

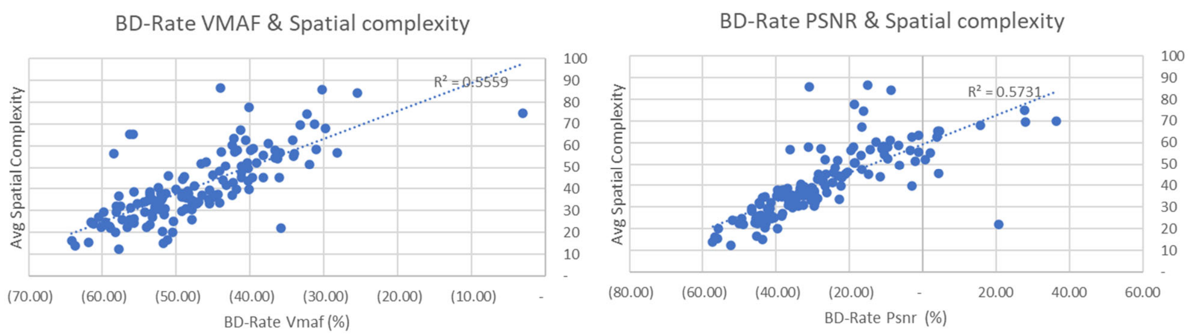


Figure 12. Correlation between spatial complexity and BD-Rates.

In the context of consistently strong BD-rate gains, LCEVC seems to outperform the native anchor less (at least on objective metrics) when more of the energy is concentrated on high frequencies. This is consistent with what is shown in 5.1, and in general with what should be expected with a multi-layer scheme: if all of the energy is on the top layer, where LCEVC cannot significantly outperform sophisticated coding schemes, coding efficiency benefits will be lower. If instead – as it is much more frequent for real-world sequences – a good portion of the energy is concentrated in mid and lower frequencies, LCEVC-enhanced encodes outperform native encodes by a greater amount.

Figure 13. Screenshots of the 5 best performing sequences

To gain an even clearer sense of the types of sequences – and relative gains – that can influence the degree of efficiency enhancement brought forward by LCEVC, we analyzed the best 5 sequences, the median 5 sequences and the worst 5 sequences in terms of BD-rate.

The best 5 sequences in terms of LCEVC BD-rate benefits achieved an average VMAF BD-Rate of −62.5% and an average PSNR BD-Rate of −54.2%. They are shown in Figure 13. These sequences are characterized by darks, talking heads, close-ups and some motion. The amount of detail varies from relatively low to relatively high. Sequences like these are typically at risk of impairments such as blocking or banding, and in fact LCEVC produces dramatic quality improvements, as exemplified with the 1080p snapshots in Figure 14.

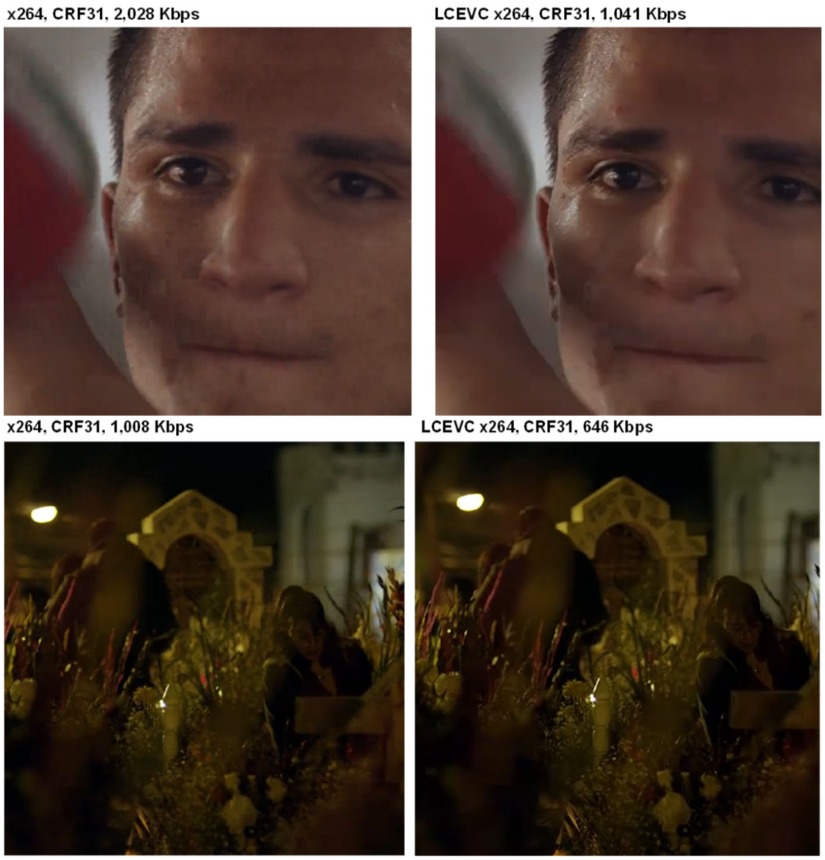


Figure 14. Snapshots of 2 exemplary sequences comparing LCEVC and x264

On these sequences, even at lower or much lower bitrate, LCEVC allows to maintain a generally lower QP in combination with a generally higher amount of details. The impact in terms of reduction of banding and "flattening" is even more evident when playing back the video, rather than analyzing a single frame.

The median 5 sequences in terms of LCEVC BD-rate benefits achieved average VMAF BD-Rate of −48.1% and average PSNR BD-Rate of −34.6%. They are shown in Figure 15a. It is difficult to characterize these sequences other than saying that they are diverse and representative sequences, including both slow and fast motion, both high and low lights, both depth of field and detail. They are indeed quite median in all senses, and we would expect them to represent the large majority of typical video content.

The worst 5 sequences in terms of LCEVC BD-rate benefits achieved average VMAF BD-Rate of −23.3% and average PSNR BD-Rate of −10.5%. They are shown in Figure 15b. These sequences, rarely a heavyweight for video transmission due to their relatively more static nature, are clearly characterized by high concentration of energy in the high frequencies. It is to be expected that benefits of LCEVC are relatively lower, since the large majority of the entropy is concentrated on the top layer, where – although LCEVC is efficient at compressing such details, as shown in 5.1 – the simplicity of LCEVC and the inherently less compressible nature of "white noisy" high frequency content (such as high contrast vegetation,

Figure 15a. Screenshots of 5 medium performing sequences

Figure 15b. Screenshots of the 5 worst performing sequences

droplets of water, etc.) does not allow for the levels of overperformance seen in other sequences.

Interestingly, despite objective metrics suggesting that LCEVC provides lower efficiency gains on these types of sequences, subjective performance of LCEVC seems to be strong, and actually similar to that of median sequences. FoodMarket, one of the sequences in the LCEVC MPEG Common Test Conditions (CTC) set, is actually among these worst 5 sequences of the El Fuente test set, but in CTC tests still managed to achieve formal MOS BD-rate gains in the range of −45% with both JM and HM.

4.2 Decoder battery consumption in real-world scenarios

On an Intel Xeon CPU E3 -1505M v5 @2.80 GHz with 32 GB of RAM, LCEVC x264 required on average 42% of the encoding time vs. native x264 encoding at full resolution, encoding 2.6 times faster in terms of frames per second.

With regards to decoding, most platforms use hardware-accelerated H.264 playback to reduce power and battery consumption while playing h.264-encoded videos. With LCEVC, decoding the H.264 base-layer is also hardware accelerated, though decoding the enhancement layer is not.

We used the following tests to compare LCEVC and native H.264 power consumption, playing back encodes with a six-core i7-9750H running Windows 10 measuring voltage and power consumed with the Open Hardware Monitor utility (https://openhardwaremonitor.org). The tests to compare power consumption involved three video files:

- Video 1 – AVC/H.264 @ 2 Mbps
- Video 2 – LCEVC AVC/H.264 @ 2 Mbps
- Video 3 – AVC/H.264 @ 4 Mbps (to roughly match LCEVC quality)

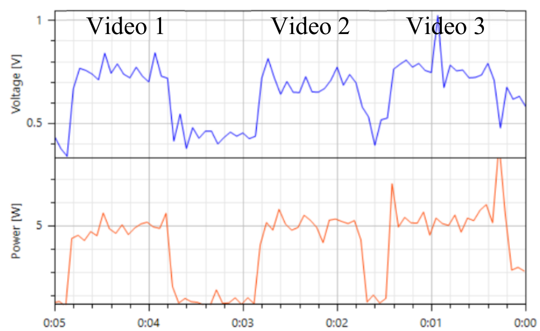


Figure 16. Decoding processing power consumption vs. native decoding.

As shown in Figure 16, LCEVC decoding consumes lower voltage and about the same power, so overall, LCEVC decode consumes less battery power. Compared to the 4 Mbps AVC/H.264 file, which is the same approximate quality as the LCEVC file, LCEVC playback is more efficient in both power and voltage. So, despite the lack of hardware acceleration for the decode of the LCEVC enhancement, LCEVC playback is substantially in line with (and actually slightly more efficient than) AVC/H.264 playback.

4.3 Encoding and decoding complexity in real-world scenarios

As also reported in [13], processing times for LCEVC were measured using the optimized implementations of LCEVC, H.264/AVC and H.265/HEVC, as described in section 4.1. The encodes and decodes have been performed on a common platform (Intel i9-8950HK_2.9GHz).

For each full resolution, the same sequences mentioned in section 4.1 were used. Table 6 reports the average timings for each resolution for both anchors and LCEVC.

Table 6. Relative encoding and decoding times for LCEVC vs. anchors (anchor = 100%)

Anchor	Resolution	EncT	DecT
H.264/AVC	HD	51.48%	96.72%
H.264/AVC	UHD	32.99%	81.88%
H.265/HEVC	UHD	34.44%	64.24%

As it can be seen, the encoding time for LCEVC is between circa 30% and 50% of the encoding time required for the anchors depending on base encoder and resolution. On the decoding side, LCEVC requires between circa 60% and 95% of the decoding time required for the anchors depending on base decoder and resolution. The low complexity of LCEVC allows power-efficient implementations of the codec via software, also at relatively high levels of the software stack. As discussed in section 2.3, the LCEVC processing is highly parallelizable due to certain characteristics of the scheme. The tools are designed to minimize the number of operations required as well as the inter-dependency between them, making efficient use of available general-purpose hardware acceleration, including SIMD, GPUs or DSPs, either alternatively or in conjunction.

5. WHY LCEVC WORKS

5.1 Coding efficiency at compressing high-frequency details

The key working principle producing LCEVC's overall compression efficiency is to suitably separate high-frequency details from mid-to-low frequency "core signal", encode the core signal at lower resolution and then efficiently encode high-frequency details with a set of very specialized tools. Although in some corner-case areas the act of encoding a lower resolution plus details may generate some "double-counting" in terms of entropy, on the whole it is understandably more entropy efficient (as we well know from "convex hull" encoding) to encode mid-to-low frequencies with a single-layer codec used at low resolution. LCEVC adds to a sort of "natively convex hull" encode the features of smart upsampling and – importantly – coding the non-predictable high-resolution details. So, part of the overall coding gain comes from the LCEVC structure, part from its signaled non-linear normative upsampling (a sort of low-complexity super-resolution) and part from its ability to efficiently code high-frequency residuals.

The key question from an entropy perspective is thus the following: how efficient is LCEVC at encoding high-frequency residuals? Can it be sufficiently efficient despite its few tools and its relatively simple temporal correlation scheme?

In order to analyze the entropy efficiency of LCEVC, we selected from the MPEG test set two 1080p source videos with high amount of spatial detail ("Cactus" and "EuroTruck") and generated difference maps between the source video and a rendition obtained by spatially downsampling the source with the LCEVC area downsampler and upsampling it to full resolution using the LCEVC "modifiedcubic" upsampling. Since the downsampled video is not compressed, this method compares the source to the asymptotic quality achievable by encoding a downsampled source ("perfect base").

We then plotted the PSNR RD curves achieved by adding to this "perfect base" an LCEVC enhancement at a range of bitrates, encoded with the LCEVC Test Model (LTM) v4.1. To validate that the LCEVC scheme is reasonably efficient at compressing this type of high-frequency residual data, we also plotted the PSNR RD curves achieved by using HM, instead of LCEVC, to encode the difference maps. In both cases, we excluded the bitrate required to encode the perfect base from our analysis, measuring only the bitrate required to encode the residuals. Notice that we know from Paragraph 3.3 that LCEVC compression performance is not very accurately "measured" by PSNR, especially in terms of cross-correlation (i.e., using PSNR to compare LCEVC with another codec), but for this comparison we went up to near-lossless quality levels, where subjective importance of differences vs. the source is less relevant and PSNR is a more accurate metric.

Admittedly, this comparison with HM is a bit unfair: LCEVC is almost two orders of magnitude less complex than HM on a per-pixel basis, the LCEVC test model has no MSE-based RDO to optimize for PSNR, and for this comparison we did not leverage all of the LCEVC coding tools (e.g., the possibility to adapt the upsampling kernel).The comparison just aims at assessing whether– despite being very low complexity – LCEVC achieves an at least reasonable compression efficiency at coding details, especially for sequences like Cactus and EuroTruck with high amounts of energy in full-resolution spatial detail (and hence not necessarily in the sweet spot for a multi-layer coding scheme).

Figure 17 shows the PSNR RD curve obtained for the Cactus and EuroTruck sequences when enhancing a perfect base, with respect to the PSNR of the perfect base upsampled with the LCEVC modified cubic upsampler.

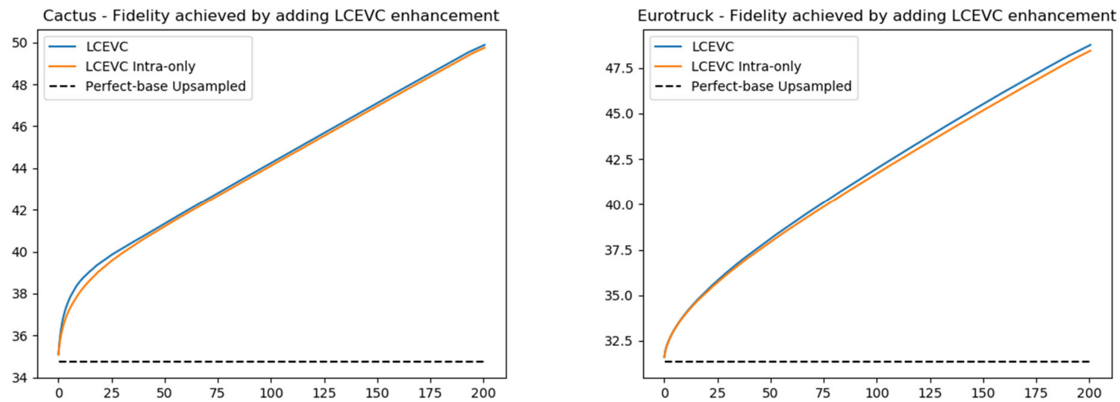


Figure 17. Cactus and EuroTruck, PSNR results based on LCEVC enhancement data

Results above show that the enhancement has significant impact on the overall performance. They also show that LCEVC is capable of reconstructing full fidelity to the original source, although it doesn't necessarily say whether LCEVC is any efficient at doing that.

We then encoded the difference between source and upsampled perfect base with HM, and then calculated the PSNR of the video obtained by adding the HM-encoded residuals to the upsampled perfect base, so as to see how LCEVC's coding efficiency compares with that.

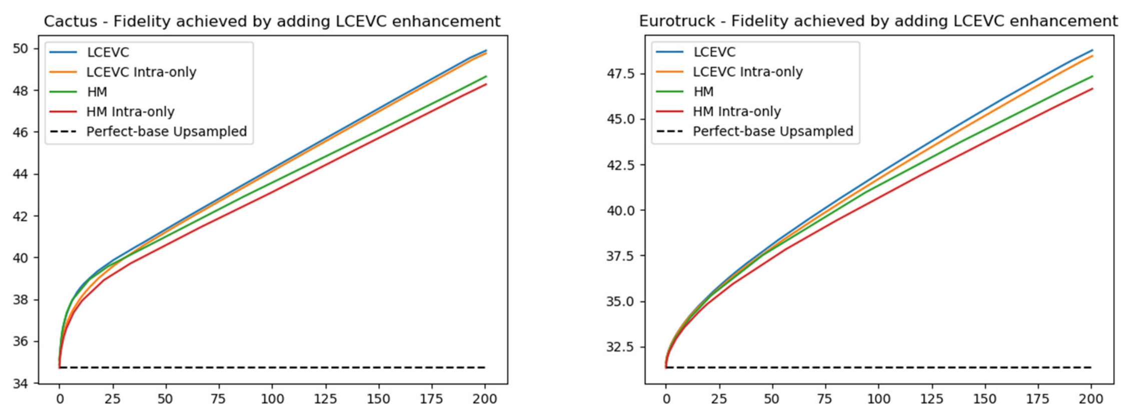


Figure 18. Cactus and EuroTruck, comparison between LCEVC and HM

Results show that, according to PSNR, LCEVC is more efficient at compressing residual high-frequency information than a much more complex coding scheme like HM.

It is also noticeable that even if the very simplified temporal scheme of LCEVC may leave ca. 0.1-0.2 dB of PSNR gain "on the table", the order of magnitude of benefit achieved from temporal correlation is similar to that of a much more complex motion-compensation-based tool set like that of HEVC. I.e., sophisticated motion-vector-based block-based motion compensation does not seem to provide substantial benefits when coding high-frequency details, especially if combined with highly efficient intra residual compression (almost 2.5 dB of PSNR gain at 50 dB between LCEVC intra-only and HM intra-only). This is likely due to the fact that high-frequency details have different depth/z-value from the camera and rarely move with uniform motion.

Lastly, Figure 18 shows that – although PSNR is not a reliable metric for assessing the coding performance of LCEVC-enhanced encodes vis a vis native encodes – at higher bitrates the PSNR performance of LCEVC seems strong. This is confirmed by looking at extended PSNR RD-curves of LCEVC-enhanced encodes vs. native encodes. In the case of Cactus,

Figure 19 below shows on the left how RD-curves look for x264 slow vs. LCEVC x264 slow when we focus on typical bitrate ranges: LCEVC-enhanced encodes provide higher PSNR scores at low bitrates and then cross-over at higher bitrates. This would seem to imply that at higher bitrates single layer coding performs better, at least in terms of PSNR. However, if we extend the bitrate range further (Figure 19, on the right), we see the LCEVC-enhanced RD-curve growing faster and eventually crossing over again, ultimately showing a gain. So, on top of the lower correlation of PSNR with formal MOS and the different slopes making PSNR unsuitable for cross-correlation for cross-codec comparisons involving LCEVC, it seems like PSNR also exhibits a wholly different behavior with LCEVC. Further analysis would be required in order to understand whether these results are “structural” to LCEVC or they are mostly due to the absence of MSE-based RDO algorithms in the current LCEVC encoder implementations.

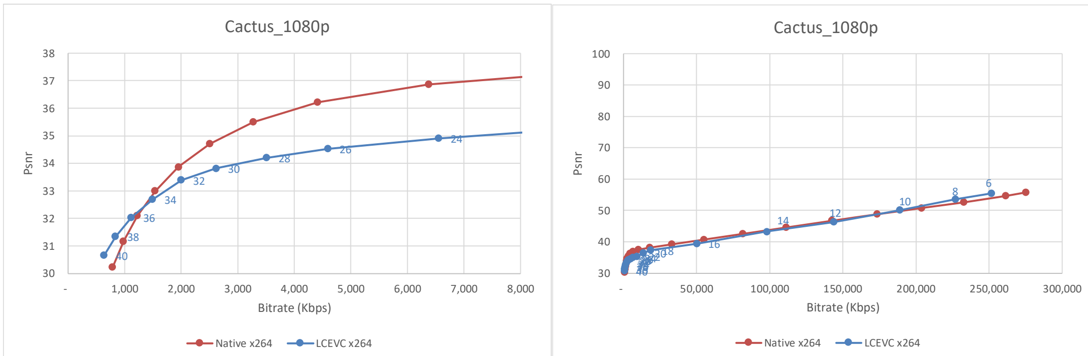


Figure 19. PSNR behavior of LCEVC encodes over normal (left) and extended (right) bitrate ranges

5.2 Accounting for Motion Consistency within Coding Efficiency

Accurate representation of motion may contribute to explain why the formal MOS performance of LCEVC is consistently superior to what measured by frame-based objective metrics (including VMAF).

This effect is particularly evident with sequences with complex motion and lots of high-frequency details, such as particularly complex eSports sequences. Figure 20 shows the VMAF and formal ITU-R BT.500 DSIS MOS RD-curves for the eSports sequences Witcher3 and Rust, with LCEVC enhancing x264 veryslow (blue) vs. native x264 veryslow (red).

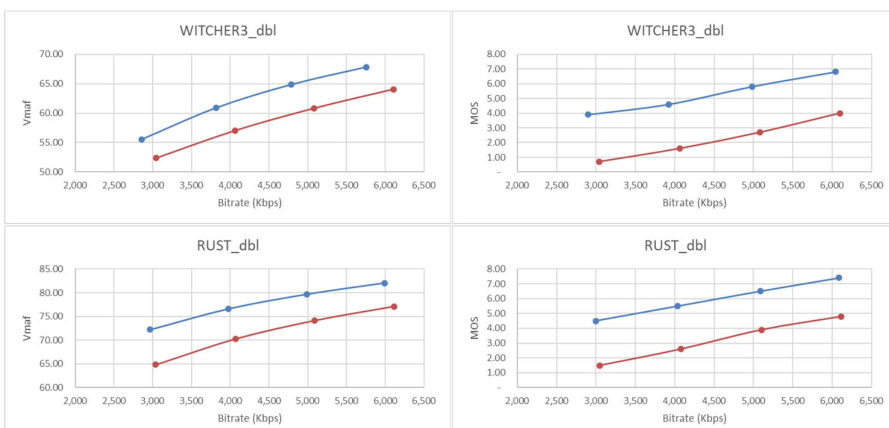


Figure 20. VMAF underestimation of MOS benefits for high-motion sequences, LCEVC x264 (blue) vs. native x264 (red).

Inspecting the encodes in order to understand the reasons for this divergence, we could observe that while frame-by-frame analysis may suggest a lower gain, watching the video in motion completely changes the assessment, due to a large amount of dragging impairments and “odd blocky motions” in native encodes, undetectable by inspecting one frame at a time.

Today's objective metrics do not measure characteristics of a "video", a three-dimensional signal in time-space, but those of individual frames of a video. As such, none of today's metrics will be able to detect how well the video is suitable for object tracking, which is what human viewers do when watching a video with moving objects.

When starved for bits, single-layer codecs degrade visual quality by rapidly sacrificing consistency of motion and object depth: a moving object becomes a flat rectangular texture that translates, often generating some degree of "dragging artifacts" with the surrounding elements, up until changes are enough to be corrected, making the object suddenly "pop" into a different texture. While these impairments may be less evident by inspecting a single frame, they become extremely evident once a viewer tries to eye-track the object and make sense of its shape. On the contrary, a softer reconstruction with high-contrast details moving properly will be eye-trackable, and the human will be able to automatically infer both the shape of the object and some of its low-contrast details, thanks to the fact that different frames sample the object in different places (the only exception being the case in which an object is static).

LCEVC inherently preserves better consistency of motion (by encoding the base at a much lower QP than would be possible with the enhanced codec used at full resolution), along with maximum coding efficiency for static details. If the overall encode is starved for bits, LCEVC will automatically sacrifice resolution of moving objects (which a human can still infer if the underlying mid-to-low frequencies are eye-trackable), and, thanks to its temporal correlation scheme, will efficiently preserve all the non-inferable details of static objects.

Although efficiency at preserving correct motion of objects across multiple frames is indeed part of compression efficiency – no frame-based objective metric is able to take it into account. Whilst formal ITU-R BT.500 DSIS MOS evaluations definitely do, and this may contribute to explain part of the divergence in Figure 20.

5.3 LCEVC vs. Smart Upsampling

Post-processing techniques for "smart upsampling" are becoming more popular, so a question may arise with regards to how LCEVC fares in the presence or as an alternative to those methods. In other words, could LCEVC-like benefits be achieved by encoding at a lower resolution and performing a smart upsample? Figure 21 below illustrates why no upsampler, however smart, can achieve similar results as LCEVC.

Lanczos upsamplers or even neural-network upsamplers perform better upsampling than traditional methods such as bicubic, although they come with considerable processing costs, by itself greater (or much greater, respectively) than the overall cost of decoding LCEVC. In addition, smart upsampling has no "failsafe": the upsampler will try to reconstruct a suitable rendition, but when it gets it wrong, there will be no way to provide the actual information necessary to reconstruct fidelity to the source content. In addition to many details being impossible to infer from a downsample, the more intelligent the AI upsampler, the more likely it is that in some cases it will "allucinate", reconstructing something similar but different (e.g., an "X" instead of a Louis Vuitton logo).

The example below in Figure 21 (magnified to better show the impact on fidelity of small details) illustrates the difference between the source input, the decoded base at half-resolution, the same decoded base upsampled with state-of-the-art Lanczos upsampling and the LCEVC-enhanced full resolution reconstruction. Notice (e.g., number "8", letter "R", letter "H") that the LCEVC-enhanced reconstruction includes high-frequency details that could not be inferred – especially for static content – from just decoding and upsampling the lower resolution encode, however smart and complex the upscaling method.

Figure 21a. Source

Figure 21b. Half-resolution decoded base

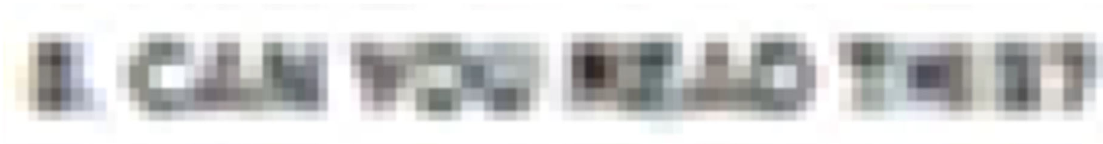


Figure 21c. Base upsampled to full resolution with FFmpeg Lanczos

Figure 21d. Full resolution LCEVC-enhanced output with typical large-scale distribution bitrates (315 Kbps of enhancement data for the whole frame, of which this caption is a small portion)

Fidelity to the source is critical to many use cases, especially whereby – for either monetization or user experience – service providers need to guarantee fidelity of advertising logos, discernability of small faces and/or readability of small text.

Thanks to residual sub-layers, LCEVC in fact combines the world of smart upsampling with the world of traditional coding: for the areas where smart upscaling is enough for high-fidelity reconstruction, LCEVC performs a low-complexity encoder-guided smart upsampling and does not need to transmit residual data; for the areas where smart upscaling fails, LCEVC allows the encoder to efficiently send the corrections that are necessary to reconstruct fidelity to the source. Importantly, LCEVC also allows the encoder to signal smart post-processing operations to be performed at the decoder, including but not limited to content-adaptive statistical dithering. As such, LCEVC could be seen as a “standard for enhanced post-processing” (including future extensibility features), with the benefit of ultimate failsafe: whenever and wherever necessary, LCEVC can efficiently – as shown in 5.1 – send the real details.

6. DISCUSSION

The results of the experiments in this contribution illustrate the following:

1. LCEVC satisfies its requirements of providing a codec-agnostic enhancement of both coding efficiency and processing complexity. As we move from AVC/h.264 to latest generation codecs, coding efficiency gains decrease whilst processing complexity gains increase, maintaining material gains across all generations.

2. LCEVC gains are confirmed – and possibly higher –in real-world scenarios with optimized implementations, across types of sequences. LCEVC encodes outperform native encoding at full resolution and provide benefits also with respect to a native encoding “convex hull”. In particular, the convex hull of LCEVC-enhanced encodes outperforms the convex hull of native encodes. Material processing benefits and suitability for light-weight software decoding are also confirmed.

3. On top of synergistic collaboration with the enhanced codec used at lower resolution (responsible for part of the coding efficiency gains), LCEVC is efficient at coding high-frequency residual data. This allows efficient reconstruction of relevant detail and capacity to reconstruct quality as close as possible to the original source, up to lossless. Efficient residual coding and fidelity to the source distinctly sets LCEVC apart from complementary approaches such as pre-processing, convex hull encoding or super-resolution.

7. CONCLUSION

LCEVC responds to the need for higher quality video in bandwidth-constrained scenarios, with an eye on sustainability of processing complexity. LCEVC design allows purely software implementation for low power hardware found on a broad range of standard devices, side-stepping the requirement for dedicated hardware blocks in the silicon. This software implementation option provides a backward-compatible capability to upgrade video services based at least in part on legacy hardware, with old and new codecs alike, whilst remaining valuable also once new codecs are fully deployed.

This contribution illustrates some of the theoretical grounds for LCEVC performance, as well as the types of performance improvements that can be expected with a statistically representative test set of sequences. Thanks to a combination of efficient coding of high frequency details and effective leverage of the enhanced codec, LCEVC satisfies its requirements and – especially for bit rates relevant to mass market video delivery – can enhance single-layer codecs to a coding performance as close as possible to that of their next generation, while maintaining compatibility with the decoder ecosystem of the enhanced codec, preserving low battery consumption for mobile devices and reducing the overall coding complexity.

ACKNOWLEDGMENT

The authors would like to thank Dr. Vittorio Baroncini, Giacomo Baroncini and GBTech laboratories (Rome-Italy) for running formal subjective assessment tests. In particular, the tests were run using special logistics in COVID-19 health safety constrains. For more information on the methodology and logistics used, please contact Dr. Vittorio Baroncini at baroncini@gmx.com.

REFERENCES

[1] Ioannis Katsavounidis "Dynamic optimizer — a perceptual video encoding optimization framework", Netflix Technology Blog.
[2] Joel Sole, Liwei Guo, Andrey Norkin, Mariana Afonso, Kyle Swanson, Anne Aaron "Performance comparison of video coding standards: an adaptive streaming perspective", Netflix Technology Blog.
[3] Jie Dong, Yan Ye, "Adaptive Downsampling for High-Definition Video Coding", IEEE Transactions on Circuits and Systems for Video Technology (2012).
[4] "Text of ISO/IEC DIS 23094-2 Low Complexity Enhancement Video Coding", ISO/IEC JTC 1/SC 29/WG 11 N19235, Alpbach, AT, April 2020.
[5] "LTM: Test model of low complexity enhancement video coding", MPEG Git, 2020. [Online]. Available: http://mpegx.int-evry.fr/software/MPEG/LCEVC/LTM
[6] K. Suehring, "H.264/AVC Software", Fraunhofer Heinrich Hertz Institute. [Online]. Available: http://iphome.hhi.de/suehring/.
[7] "High Efficiency Video Coding (HEVC)", Fraunhofer Heinrich Hertz Institute. [Online]. Available: https://hevc.hhi.fraunhofer.de/.
[8] "Versatile Video Coding (VVC)"
[9] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara (2016), "Toward A Practical Perceptual Video Quality Metric", Netflix Technology Blog, 06 Jun. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652.
[10] Z. Li et al. (2018), "VMAF: The Journey Continues", Netflix Technology Blog, 25 Oct. [Online]. Available: https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12.
[11] Rec. ITU-R BT 500-14 "Methodologies for the subjective assessment of the quality of television images".
[12] G. Bjøntegaard (2001), "Calculation of average PSNR differences between RD-curves", ITU-T VCEG-M33
[13] F. Maurer, S. Battista, L. Ciccarelli, G. Meardi, S. Ferrara "Overview of MPEG-5 Part 2 – Low Complexity Enhancement Video Coding (LCEVC)", ITU Journal: ICT Discoveries, Vol. 3(1), 8 June 2020.