SCTE CABLE-TEC EXPO 22
SEPTEMBER 19-22 • PHILADELPHIA, PA

2022 Fall Technical Forum
SCTE • CABLELABS • NCTA

Perceptual Video Coding Optimization Techniques: Most Recent Trends and Future Directions

A Technical Paper prepared for SCTE by

Dan Grois, PhD
Principal Researcher
Comcast
Comcast Center 1701 JFK Blvd. Philadelphia, PA 19103
1-215-286-1700
dan_grois@comcast.com

Alex Giladi
Fellow
Comcast
Comcast Center 1701 JFK Blvd. Philadelphia, PA 19103
1-215-286-1700
alex_giladi@comcast.com

Table of Contents

List of Figures

List of Tables

1. Introduction

There is currently a strong demand for high resolution video content, particularly for the high-definition (HD) and UltraHD video content to be displayed on a variety of devices, ranging from Smart TVs and laptops to mobile devices and smartwatches. There is a continuous need to decrease video transmission bit-rate, especially for delivery over wired and/or wireless/cellular networks without reducing visual presentation quality [1]-[4], [5].

In addition, the HDR UltraHD video content is recently attracting a lot of attention due the relatively high luminance levels and fine shadow details, which extend much beyond conventional Standard Dynamic Range (SDR) content. The HDR technology makes it possible to present highly bright signals along with very dark signals on the same video frame, thereby providing a high contrast ratio within the same image. In addition, the HDR video content is usually combined with a Wide Color Gamut (WCG), such as BT.2020, thereby enabling to present video with a significantly extended color spectrum. Particularly, HDR has gained its popularity after the development and approval of the High Efficiency Video Coding (HEVC) standard, i.e. H.265/MPEG-HEVC, in 2013 [6].

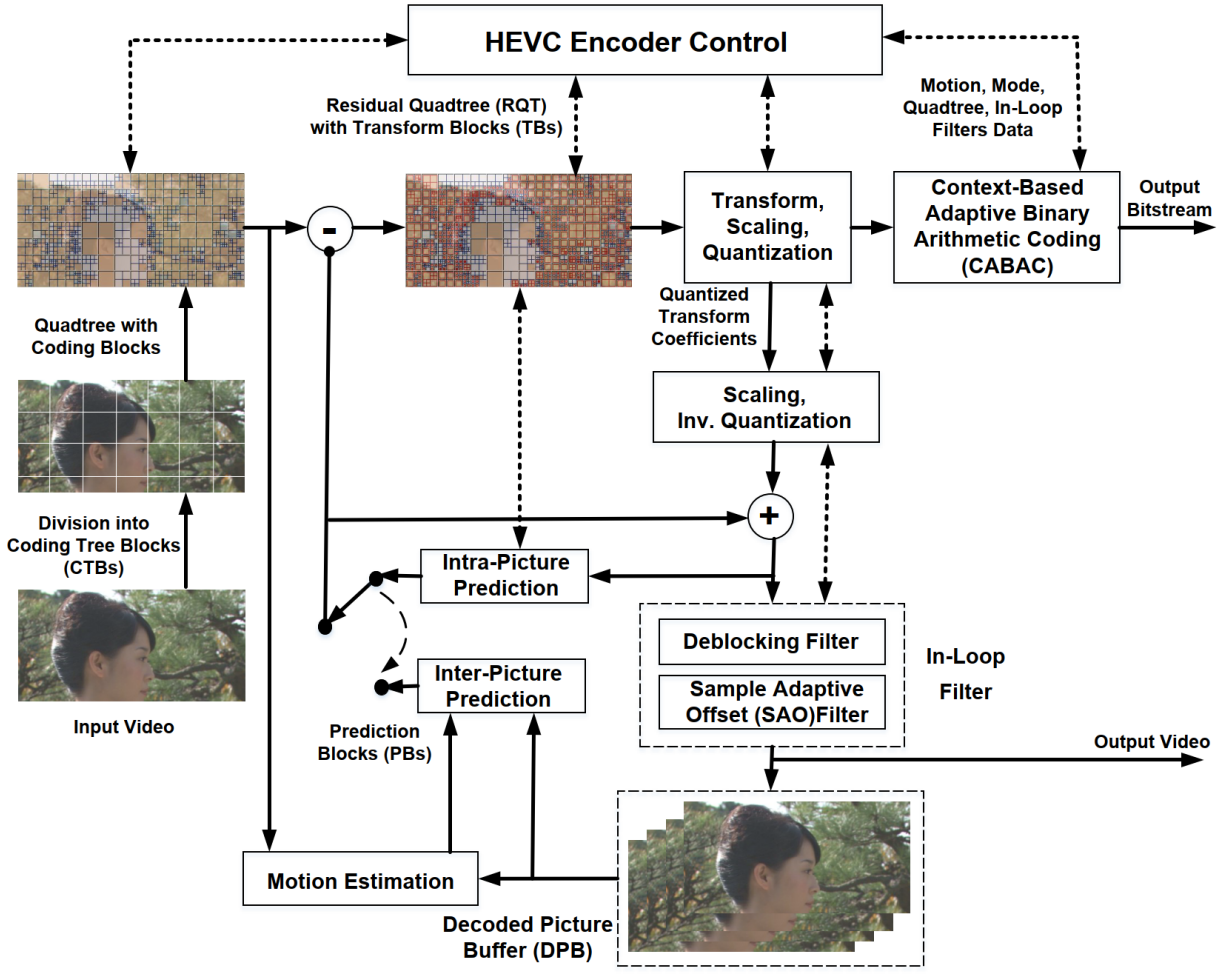

Figure 1 – The schematic block diagram of the H.265/MPEG-HEVC encoder.

The development of the first version of HEVC by the Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Pictures Expert Group (MPEG) was officially finalized in January 2013 [6]. *Figure 1* illustrates a block diagram of the H.265/MPEG-HEVC encoder. After that, the final aligned HEVC specification was approved by ITU-T as Recommendation H.265 and by ISO/IEC as MPEG-H, Part 2. About one year later, the 2nd HEVC version was finalized, incorporating the Range Extensions (REext) as well as the Scalable and Multi-view Extensions (SHVC and MV-HEVC, respectively) [7]. In turn, the 3rd and 4th HEVC edition were issued in 2015 and 2016, further containing the 3D Video Coding Extensions (3D-HEVC) and the Screen Content Coding Extension (HEVC-SCC), respectively [8], [9]. When developing the H.265/MPEG-HEVC standard, high-resolution video coding was considered as one of its main potential application scenarios, while keeping it applicable to almost all existing use cases that were already targeted by H.264/MPEG-AVC. The development process of H.265/MPEG-HEVC was driven by the most recent scientific and technological achievements in the video coding field. As a result, when compared to its predecessor - H.264/MPEG-AVC, H.265/MPEG-HEVC is able to achieve a bitrate reduction of roughly 50% for substantially the same visual quality [1]-[4].

Video applications continue to gain a lot of traction and to have an enormous demand. A very significant increase in the bandwidth requirements is expected by 2023, particularly due to the increase in the resolution supported by devices. It is expected that 66% of the connected flat-panel TV sets will have the support for the Ultra-High Definition (UltraHD) resolution compared to only 33% in 2018 (note that "UltraHD" in this paper refers to the 3840x2160 resolution, also known as 4K or 2160p). The typical bitrate for a 60fps 4K HDR10 video is between 15 to 24 Mbps, nearly four times the typical High-Definition (HD) video bitrate [5].

As a result, there is a continuous strong need to further decrease video transmission bitrate, especially for the UltraHD content, substantially without reducing the perceptual visual quality.

2. Background: Human Visual System

One the most popular approaches for improving video quality is related to considering *spatial frequency sensitivity* of Human Visual System (HVS). As known, the HVS system is a part of the central nervous system, which enables processing of visual details and generating non-image photo response functions by obtaining and processing visible information. Thus, for example, during the coding process, the values of the Discrete Cosine Transform (DCT) frequency coefficients can be attenuated by applying quantization matrices: i.e. lower spatial frequencies are usually quantized with smaller quantization parameters (QPs), while higher spatial frequencies - with larger QPs [10], [11].

However, the improvements in visual quality of the related state-of-the-art approaches are relatively small, and more efficient solutions are desirable. In addition, most of the state-of-the-art pre-processing methods are designed for the relatively low-resolution SDR video content, and as a result, these methods found to be mostly inefficient for HDR UltraHD.

In turn, this is also true for the coding schemes that aim to remove fine details below a predefined visibility threshold, which is referred as Just Noticeable Difference (JND). As a result, the state-of-the-art JND-based schemes do not provide sufficient video quality improvement for the HDR UltraHD video content as well.

In this paper, two important perceptual video coding optimizations techniques are presented and discussed in details: *Section 3* describes a novel perceptual video optimization framework based on authors' work presented in [10], [11], and *Section 4* describes a joint backward and forward masking encoding scheme, based on authors work presented in [12]. Then, future directions for perceptual video coding optimizations are provided in *Section 5*, while this paper is concluded in *Section 6*.

3. Perceptual Video Quantization Framework

The human visual system (HVS) is considered to be a very complex system, while a level of contrast that is required to generate a response perceived by HVS is known as a contrast threshold of a sinusoidal luminance pattern. In turn, an inverse of this threshold is called "contrast sensitivity", and it varies as a function of a spatial frequency.

3.1. Background: Contrast Sensitivity Function

The relationship between the spatial frequency and contrast sensitivity is known as a contrast sensitivity function (CSF) that differs for achromatic and chromatic scenes. The term Contrast Sensitivity (CS) often relates to visual acuity, thereby being able to differentiate between the object and the background [10].

In turn, CSF generally defines the sensitivity of the observer to various frequencies of visual stimuli, e.g., sensitivity to vertical black and white strips grating as a function of spatial frequencies [13],[14]. In case, the above frequencies are higher than a threshold predefined by the Human Visual System (HVS), the human observers are not able to differentiate between the strips. Also, generally, the HVS sensitivity to luminance significantly differs from the HVS sensitivity to chrominance.

The HVS is more sensitive to low spatial frequencies than to high spatial frequencies [15]-[18], and by assuming that HVS is isotropic, it can be modeled as a nonlinear point transformation that is followed by a Modulation Transfer Function (MTF) [19].

Later, this approach was practically used in developing a HVS-based quantization table for the JPEG still image compression standard [20], [21]. Authors of [21] derived this table by incorporating a HVS model developed by Daly [15]-[18] with an uniform quantizer, and further claiming that by replacing the JPEG quantization table with their HVS-based quantization table, obvious perceptual quality improvements are achieved. More specifically, the authors of [21] applied a 1st order low-contrast MTF of the HVS model proposed by Daly for generating a HVS-based quantization table for the baseline JPEG image compression standard, as follows below.

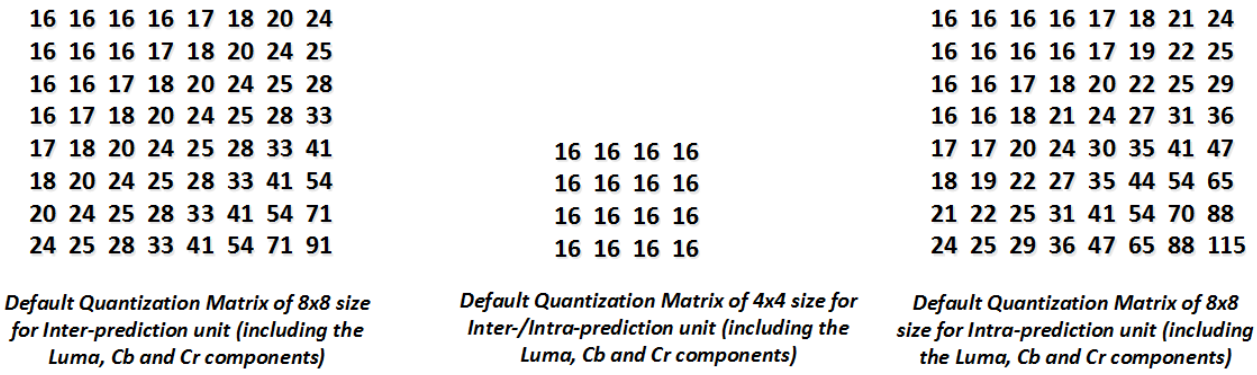

Figure 2 – The default perceptual quantization matrices defined in the HEVC standard.

The HEVC video coding standard allows usage of perceptually-tuned frequency-dependent quantization matrices, instead of applying a constant quantization parameter (QP) on each coding block. These matrices better suit the HVS characteristics by allowing to quantize higher frequencies in a stronger manner, while their sizes vary from 4x4 to 32x32. However, the specification of the HEVC standard [6]-[9] only defines default quantization matrices for 4x4 and 8x8 transform blocks (see *Figure 2*).

The rest of matrices, i.e. for transform block sizes of 16×16 and 32×32, are obtained by upsampling the original 8×8 perceptual quantization matrix respectively. More specifically, the original 8×8 matrix is replicated: each block in the 8×8 matrix is replicated to the 2×2 area of the 16×16 transform block and to the 4×4 area of the 32×32 transform block, as shown in *Figure 3*.

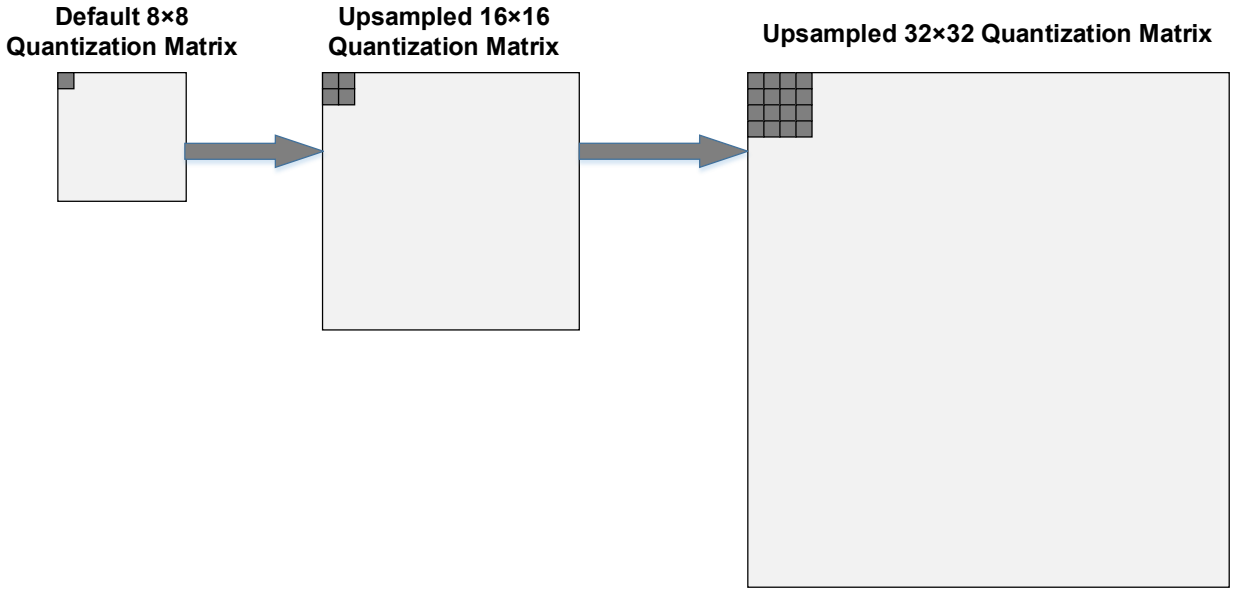

Figure 3 – Upsampling the default 8×8 HEVC matrix for obtaining default matrices for 16×16 and 32×32 transform block sizes.

Depending on the transform block type (i.e. used for *Intra* or *Inter*-picture prediction) and transform block size (i.e. 4×4, 8×8, 16×16 or 32×32), the HEVC standard employs twenty quantization matrices: 8 matrices for *Y* (*Luma*) component and 6 matrices for each of *Cb* and *Cr* (*Chroma*) components, as specified in *Table 1* below.

Table 1 - HEVC Transform Block Type/Size-Dependent Quantization Matrices.

Block Component	Type/size-Dependent Quantization Matrices
Y (*Luma*)	Intra 4×4, Intra 8×8, Intra 16×16, Intra 32×32; Inter 4×4, Inter 8×8, Inter 16×16, Inter 32×32.
Cb (*Chroma*)	Intra 4×4, Intra 8×8, Intra 16×16; Inter 4×4, Inter 8×8, Inter 16×16.
Cr (*Chroma*)	Intra 4×4, Intra 8×8, Intra 16×16; Inter 4×4, Inter 8×8, Inter 16×16.

In addition, HEVC allows to use other quantization matrix values (i.e. *customized* quantization matrix values) besides the default values. For that, the above-mentioned customized quantization matrix values can be transmitted within the HEVC bitstream Sequence Parameter Set (SPS) or Picture Parameter Set (PPS), while coding these customized values by using so called Differential Pulse Code Modulation or in short DPCM. Similarly, the 16×16 and 32×32 quantization matrices are obtained by upsampling corresponding 4×4 and 8×8 quantization matrices (see *Figure 3*).

In spite of the fact that the HEVC default perceptual quantization matrices of *Figure 3* are based on HVS, they were initially developed and tested on low-resolution JPEG images, such as 512×512 pixels. Therefore, they almost didn't provide any benefits for UltraHD video content, that has the 3840x2160 resolution in terms of luma samples, which is the most popular resolution nowadays. As a result, this is currently also a reason for the relatively low popularity of these default perceptual quantization matrices, which most often are not used at all.

In the following section, the design and development of novel perceptual quantitation matrices for encoding UltraHD HDR video content is presented, further being inspired by investigating CSF of a human visual system. The novel perceptual quantitation matrices significantly improve perceived video quality without a need for pre-processing and without an increase in coding computational complexity.

3.2. Perceptual Quantization Matrices for UltraHD Resolution Displays

The contrast sensitivity of the human eyes, and more generally – of the human visual system as the whole, is one of the main factors how humans perceive achromatic or chromatic images. Therefore, when developing HVS-based models, it is especially important to determine the Contrast Sensitivity Function (CSF) as accurate as possible. For that, it is important to consider substantially all known HVS characteristics that have any impact of the CSF [22].

With this regard, in addition to the HVS-based models developed by Mannos&Sakrison and Daly at the end of the 20th century, Barten in his paper from 2004 proposes a more accurate HVS-based physical model/formula for the contrast sensitivity of the human eye. Particularly, in his work, Barten considers a plurality of HVS parameters, such as photon noise, neural noise, external noise, lateral inhibition, eye pupil diameter, eye pupil size, angular size of the object, luminance conditions, etc.

As a result, Barten's HVS-based CSF model is considered to be the most accurate for representing the HVS contrast sensitivity, and considered to be the best CSF model to date [23]-[25]. However, Barten's model is very complex, and its usage for an accurate determining of efficient perceptual quantization matrices, either in HEVC or in other emerging video coding standards, is found to be very challenging.

Therefore, at the 1st step of authors work [10], perceptual quantization matrices to be employed during the video coding loop have been designed by fitting the Daly's HVS-based model into the Barten's HVS-based CSF model.

In turn, at the 2nd step, the CSF-tuned human visual coefficients are empirically optimized by gradually attenuating high frequencies in a much stronger manner than low frequencies, and further giving priority to luminance (Luma) over chrominance (Chroma) due to the fact that human eye is more sensitive to Luma changes than that of Chroma [10].

For obtaining experimental results presented in this section, a special emphasis was made on video sequences having a 10-bit sample representation and UltraHD spatial resolution (particularly, the 4K resolution – i.e. 2160p, or more specifically, the 3840x2160 resolution in terms of luma samples), as presented in *Table 2* below.

Table 2 - HDR UltraHD test video sequences.

Tested Video Sequences	No. of Frames	Frame Rate per Second	Resolution	Dynamic Range
Lucy (provided by NBCUniversal®)	8425	24	3840x2160	HDR
Everest (provided by NBCUniversal®)	7202	23.98	3840x2160	HDR
Warcraft (provided by NBCUniversal®)	8177	23.98	3840x2160	HDR
Regatta (provided by UltraHD forum®)	5841	59.94	3840x2160	HDR

In *Figure 4* below, sample frames of the above-mentioned tested sequences are presented.

(a) (b)

(c) (d)

Figure 4 – A sample frames from the tested sequences: (a) "Lucy"; (b) "Everest"; (c) "Warcraft"; (d) "Regatta".

These tested video sequences can be generally characterized as follows:

- "Lucy" – includes many action scenes, many fast motion scenes, mixed content [26];
- "Everest" – includes mountains views, many snow scenes, mostly slow motion scenes [26];
- "Warcraft" – includes various computer-generated content, mostly fast motion scenes [26];
- "Regatta" – includes many water scenes, many fast motion scenes [27].

The x265 open source HEVC-based encoder [28] was selected for implementing the proposed perceptual quantization framework due its ubiquity in the industry and flexibility in configuration, as well as due to its

good coding performance. When encoding the video sequences of *Table 2* with the target bit-rates of 8Mb, 10Mb, 12Mb and 14Mb, the BD-BR coding gains are significantly large in terms of both PSNR [29] and SSIMPlus [30], [31]. Specifically, the BD-BR PSNR and BD-BR SSIMPlus bit-rate savings for the HEVC encoding with the proposed QMs versus HEVC encoding with the constant QP reach coding gain of 15.5%, for encoding the "Lucy" video sequence, as is shown in *Table 3 below*.

Table 3 - BD-BR PSNR and SSIMPlus bit-rate savings for the HEVC encoding.

Tested Video Sequences	BD-BR SSIMPlus Proposed QMs vs. Default HEVC QMs	BD-BR SSIMPlus Proposed QMs vs. no QMs
Lucy	-13.5%	-15.5%
Everest	-3.8%	-5.9%
Warcraft	-6.1%	-6.0%
Regatta	-10.6%	-11.9%

As can be clearly seen from *Table 3*, by employing the proposed perceptual QMs, significant coding gains of up to about 16% are achieved.

Especially, these coding gains are significant in terms of the SSIMPlus metrics, while for the "Regatta" video sequence the coding gain is the most significant - the "Regatta" video content is considered to be hard to encode, since it contains many water scenes, and the proposed perceptual QMs perform much better for such content.

In turn, *Table 4* below presents BD-BR PSNR and BD-BR SSIMPlus bit-rate savings for the HEVC encoding with the proposed perceptual QMs versus HEVC encoding with the constant QP per Common Test Conditions (CTC) defined in [32] – i.e. without employing the default HEVC QMs.

Table 4 - BD-BR PSNR and SSIMPlus bit-rate savings for the HEVC encoding.

Tested Video Sequences	BD-BR PSNR Proposed QMs vs. no QMs	BD-BR SSIMPlus Proposed QMs vs. no QMs
Lucy	-2.5%	-6.3%
Everest	-0.9%	-5.8%
Warcraft	-1.0%	-7.9%
Regatta	-2.4%	-11.3%

In this case, as seen from *Table 4*, the coding gain as a result of employing the proposed QMs is even larger and is up to 11.3%.

Below, as an example, the breakdown of the "Regatta" video sequence is presented, thereby showing the SSIMPLus score in a range between *0* and *100*, while the larger the number - the better the video quality is (*100* is the best possible quality). As is clearly seen from *Table 5*, when the proposed perceptual QMs are employed, the SSIMPlus score [30], [31] is significantly higher – i.e. it is more than *1 point* - compared to encoding with a constant QP per CTC [32] (i.e. marked as "no QMs" in the above table). Similarly, the encoding with the default HEVC QMs provides a little improved visual quality compared to the above-mentioned constant QP encoding, but still much worse quality compared to the encoding with the proposed perceptual QMs. In turn, the minimal SSIMPlus score in increased by *1 point* for the bit-rates of 8Mb, 10Mb, 14Mb, and by even *2 points* for the bit-rates of 6Mb, 12Mb, which is visually noticeable.

Table 5 - SSIMPlus scores for encoding the Regatta video sequence.

Target Bit Rate (Kb)	SSIMPlus (no QMs)	SSIMPlus (Default HEVC QMs)	SSIMPlus (Proposed QMs)	Minimal SSIMPlus (no QMs)	Minimal SSIMPlus (Default HEVC QMs)	Minimal SSIMPlus (Proposed QMs)
6,000	83.09	83.14	84.16	68	68	70
8,000	86.31	86.42	87.49	74	74	75
10,000	88.61	88.74	89.80	78	78	79
12,000	90.28	90.42	91.40	81	81	83
14,000	91.49	91.65	92.56	84	84	85

Perceptual quantization matrices for mobile devices, such as tablets and smartphones, which have much smaller display sizes, thereby allowing a significant reduction in the overall video transmission bit-rate by removing non-perceivable details from each video frame, are discussed at the next section below.

3.3. Perceptual Quantization Matrices for Mobile Device Displays

For mobile devices, which have smaller display sizes, there is a need to develop dedicated perceptual quantization matrices for coding High Dynamic Range (HDR) mobile device-based video content. The perceptual quantization matrices proposed at [11] are based on Human Visual System (HVS) and utilized for reducing video transmission bit-rate and for optimizing perceived visual quality of video content to be displayed on mobile devices, such as tablets and smartphones.

According to video coding scheme proposed at [11], visual quality of the HDR UltraHD video content is significantly improved, for substantially the same bit-rate, in terms of the popular objective quality metric SSIMPlus. On the other hand, the video transmission bit-rate is significantly reduced by up to about 25%, while keeping visual quality of the video content, to be displayed on a mobile device screen, substantially at the same level.

Similarly to what is explained at the previous section, the x265 open source HEVC-based encoder [28] was selected for implementing the proposed perceptual quantization framework due its ubiquity in the industry and flexibility in configuration, as well as due to its good coding performance.

Table 6 below presents, in its right column, the BD-BR SSIMPlus bit-rate savings for the HEVC encoding with the proposed perceptual quantization matrices (QMs) that are optimized for *mobile devices* versus HEVC encoding with the default QMs, as defined in the HEVC specification. In addition, in the middle column, are presented the BD-BR SSIMPlus bit-rate savings for the HEVC encoding with the proposed perceptual QMs that are optimized for *mobile devices* versus HEVC encoding with the constant QP as defined in CTC [32]– i.e. without employing the default HEVC QMs. As can be clearly seen from *Table 6*, by employing the proposed perceptual QMs, significant coding gains of up to about 25% are achieved. It should be noted that for the "Regatta" video sequence the coding gain is the most significant - the "Regatta" video content is considered to be hard to encode, since it contains many water scenes, and the proposed perceptual QMs perform much better for such content. Below, as an example, the quality scores for encoding the "Regatta" video sequence with target bit rates varying between 2Mb and 5Mb are presented, thereby showing the SSIMPLus score in a range between *0* and *100*, while the larger the number - the better the video quality is (*100* is the best possible quality).

Table 6 - BD-BR SSIMPlus bit-rate savings for the HEVC encoding.

	BD-BR SSIMPlus	BD-BR SSIMPlus

Tested Video Sequences	Proposed QMs vs. Default HEVC QMs	Proposed QMs vs. no QMs
Lucy	-15.5%	**-16.8%**
Everest	-21.4%	**-22.2%**
Warcraft	-5.9%	**-7.8%**
Regatta	-22.2%	**-23.9%**

Also, as it is clearly seen from *Table 7*, when the proposed perceptual QMs that are optimized for *mobile devices* are employed, the SSIMPlus score is significantly higher – i.e. it is up to about 2 *points* compared to encoding with a constant QP according to CTC [32], i.e. marked as "no QMs". Similarly, the encoding with the default HEVC QMs provides a little improved visual quality compared to the above-mentioned constant QP encoding, but still much worse quality compared to the encoding with the proposed perceptual QMs. In addition, the minimal SSIMPlus score in increased by a very significant number of up to *7 points* for the bit-rate of 3Mb, which is visually clearly noticeable.

Table 7 - SSIMPlus scores for encoding the Regatta video sequence.

Target Bit Rate (Kb)	SSIMPlus (no QMs)	SSIMPlus (Default HEVC QMs)	SSIMPlus (Proposed QMs)	Minimal SSIMPlus (no QMs)	Minimal SSIMPlus (Default HEVC QMs)	Minimal SSIMPlus (Proposed QMs)
2,000	76.82	76.85	**78.61**	48	48	**54**
3,000	82.06	82.10	**83.48**	58	58	**65**
4,000	84.82	84.87	**86.10**	65	66	**72**
5,000	87.20	87.28	**88.48**	73	73	**77**

In the next section, another promising approach for increasing video coding gain is discussed. In this approach, "visual masking" is applied [12], which is based on the human visual system (HVS) characteristics.

4. Perceptual Video Masking Framework

As known, HEVC was especially designed for coding of HD and UltraHD video content with a much larger coding gain compared to its predecessor H.264/MPEG-AVC, thereby reducing both spatial and temporal video content redundancies in a much more efficient way, which in turn significantly assisted in compression of the HDR UltraHD video content [1]-[5]. However, coding of the HDR video content still remains challenging due to users' demands for high visual quality, which in turn requires allocating more bits and increasing a video coding depth (e.g., from 8 bits to 10 bits). In addition, the transmission bandwidth is normally limited due to a typical limitation of the existing network infrastructure, especially in case of the transmission over wireless/cellular networks. As a result, in order to stay within the transmission bandwidth limits, the high-resolution HDR video content is often compressed with visually perceived coding artifacts. Moreover, encoding of the HDR content normally consumes significant computational resources due to a requirement to preserve fine details within the HDR video. Therefore, there is further a strong demand to improve perceived visual quality of the compressed HDR video substantially without increasing its bit-rate [10], [11].

4.1. Background: Visual Masking

One of the promising approaches for increasing video coding gain is applying "visual masking", which is based on a very interesting phenomenon observed in the human visual system (HVS) [12]. According to this phenomenon, two or more stimuli are presented sequentially to a viewer, with one stimulus acts as a target which has to be detected and described, while other stimuli are used to mask the visibility of that target. With this regard, a good amount of research has been carried out in the video compression field, such as [33],[34] for example, which exploits the above-mentioned phenomenon by providing a psycho-visual algorithm that has been implemented in the x264 encoder [35]. In turn, more advanced studies of are further presented and discussed in [36]. In addition, in the most recent work, such as [37], it is proposed to mask temporal activities that are unnoticeable by human visual system by using a masking coefficient. Further, [38] presents a video Just Noticeable Difference (JND) scheme by employing compound spatial and structure-based temporal masking, further measuring a JND threshold for each transform coefficient of a color video. Also, [39] proposes an improved transform-based JND estimation model considering multiple masking effects.

However, all surveyed existing visual masking approaches, the most interesting of which are indicated above, lead to relatively low bitrate savings. As a result, these approaches have not been adopted in the video streaming/coding industry to date. In addition, computational complexity of existing visual masking schemes is relatively high due to the utilization of relatively complex quantization models [19].

4.2. Forward and Backward Masking Encoding Scheme

In this section, a masking technique for videos is exploited and discussed in detail. Extensive experiments have been carried out for the unidirectional (either forward or backward) temporal masking, but due to the lower coding gains when compared to the bidirectional (i.e. *joint* forward and backward) temporal masking, and to keep the presentation of the experimental results of this work in a clear and simple manner, this paper is focused on the bidirectional temporal masking only. The x265 open source HEVC-based encoder [28] was selected for implementing the proposed joint forward and backward temporal masking framework due its ubiquity in the industry and flexibility in configuration, as well as due to its good coding performance.

With this regard, *Figure 5* presents a schematic illustration of the proposed framework, which includes three forward sub-windows *1* to *3*, and three backward sub-windows *4* to *6*. Each window can have a different length and for each window, a set of different quantization parameters (QPs) can be assigned by adding the following QP offsets: ΔRef_1, $\Delta Non\text{-}Ref_1$; ΔRef_2, $\Delta Non\text{-}Ref_2$; ΔRef_3, $\Delta Non\text{-}Ref_3$; ΔRef_4, $\Delta Non\text{-}Ref_4$; ΔRef_5, $\Delta Non\text{-}Ref_5$; ΔRef_6, $\Delta Non\text{-}Ref_6$. Also, different QP offsets can be assigned separately to reference (e.g., *P-frames*, *B-frames*) frames and to non-reference frames (e.g., *b-frames*) present inside each masking window. The above-mentioned QP offsets are predefined in the x265 code [28] for reference *B-frames* and for non-reference *b-frames*, the offsets for *P-frames* are automatically reduced by 30%, thereby applying only 70% of the ΔRef offset value, to improve their quality and to increase a coding gain. In addition, no QP offsets are applied to *I frames*, regardless of the fact whether the *I-frame* is a scenecut or not. In case when an *I-frame* is present inside a masking window, the masking is avoided on all frames after this *I-frame* in a given masking direction (either forward or backward). The values of QP offsets can be customized using the x265 command line [28].

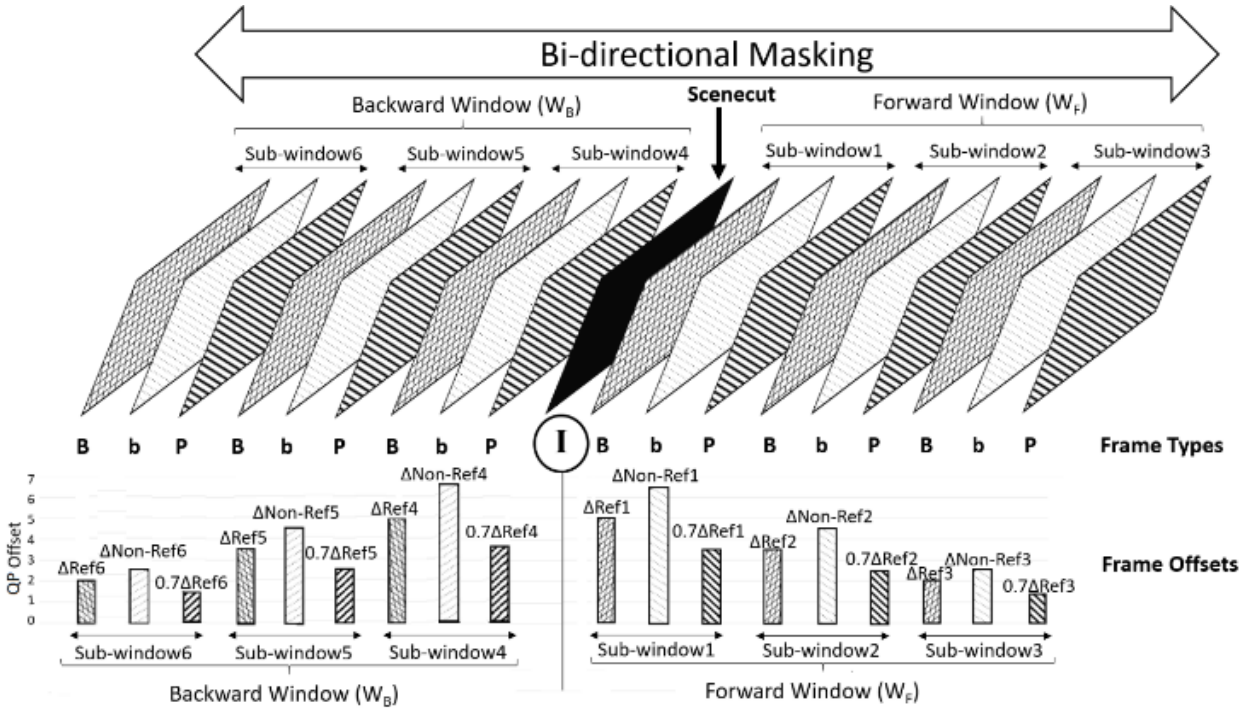

Figure 5 – A schematic illustration of the proposed joint backward and forward temporal masking framework.

For evaluating the proposed framework, the authors selected a wide range of cinematic content, mostly in 10-bit UltraHD resolution, which can be characterized as follows: (a) "El Fuente" – includes mixed content, with both fast and slow motion; (b) "Lucy" – includes many action scenes, many fast motion scenes, mixed content; (c) "Warcraft" – includes various computer-generated content, mostly fast motion scenes; (d) "Everest" – includes mountains views, snow scenes, mostly slow motion scenes; (e) "Regatta" – includes water scenes, many fast motion scenes. Main technical parameters of the above-mentioned cinematic content are presented in *Table 8* below:

Table 8 – Test Sequences.

No	Sequence name	Resolution	Frame count	Frame rate	Duration (sec.)	Bit depth
1	**El Fuente**	3840x2160	1500	60	25	8
2	**Lucy**	3840x2160	480	24	20	10
3	**Warcraft**	3840x2160	495	23.98	~20	10
4	**Everest**	3840x2160	480	23.98	~20	10
5	**Regatta**	3840x2160	1199	59.94	~20	10

The test environment, lighting, and the rest of requirements for obtaining optimal viewing conditions were set according to [40].

4.2.1. Experimental Results and Brief Discussion

The experiments were conducted according to [40],[41]. Due to the COVID-19 restrictions, viewing sessions were done remotely, and strict instructions were provided to all participants in accordance with [40],[41]. The video playback was done on high-end consumer 4K TV displays capable of playing HEVC-encoded video content with a minimal screen size of 55'' (OLED TV). High-quality 32'' professional SDI reference/grading displays were used as well.

Table 9 – Bitrate Savings in Terms of BD-BR.

Sequence Name	CRF	Without Masking (Reference)		With Masking (Tested)		BD-BR Savings
		Bitrate	MOS	Bitrate	MOS	
El-Fuente	20	23960.64	88	18266.86	89	-10.7%
	24	14701.49	85	11378.59	82	
	28	9013.59	80	7115.96	79	
	32	5629.3	73	4560.42	72	
	36	3616.78	64	3054.72	64	
Lucy	20	14074.81	89	11390.49	88	-5.6%
	24	8042.85	85	6677.24	84	
	28	5003.31	74	4210.82	75	
	32	3211.48	69	2737.33	62	
	36	2113.21	63	1832.32	55	
Warcraft	20	7092.24	88	6655.14	88	-2.3%
	24	4082.99	85	3852.15	85	
	28	2572.63	82	2436.43	83	
	32	1705.72	73	1621.98	71	
	36	1171.32	66	1121.69	58	
Everest	20	13420.24	85	11418.12	85	-14.8%
	24	5738.53	83	4943.96	82	
	28	2516.55	76	2224.67	78	
	32	1480.22	69	1335.23	70	
	36	1007.73	65	922.89	65	
Regatta	20	34769.29	83	26042.09	84	-26.3%
	24	21264.06	80	15836.46	82	
	28	13164.48	82	9791.08	79	
	32	8342.09	76	6249.4	72	
	36	5349.47	66	4192.14	61	
Average	-11.9%					

As seen from *Table 9*, the largest bitrate savings of more than 26% are for "Regatta" sequence, which has the highest bitrate, on average.

On the other hand, the smallest bitrate savings of 2.3% are for the "Warcraft" sequence, which has the smallest bitrate, on average. Further, there is a substantial decrease in the overall computational complexity in terms of encoding times (for simplicity, the results are not presented).

One of the *important findings* is that the proposed joint backward and forward temporal masking framework tends *to perform better for higher bitrates and frame rates,* as well as for content that includes textures, such as water and snow.

In this paper, two approaches have been presented and discussed in detail: perceptual quantization matrices and visual masking. Future directions with this regard are discussed at the next section.

5. Future Directions for Perceptual Video Coding Optimizations

In spite of the fact that the HEVC standard was especially designed for the HD and UltraHD video content, more efficient video compression techniques are still desired, especially for streaming UltraHD video content as well as Panorama video content (so called 360° video content) from concerts, shows, sport events, etc. Therefore, in order to fulfill this demand, the exploration phase for future video coding technologies beyond HEVC (ITU-T H.265 | ISO/IEC 23008-2) started in October 2015 by establishing a Joint Video Exploration Team (JVET) on Future Video Coding of ITU-T VCEG and ISO/IEC MPEG. In turn, these future technologies were integrated into the Joint Exploration Test Model (JEM), and the official standardization activities for the next-generation video coding standard officially started in April, 2018 - right upon publishing results of the Call for Proposals (CfP) for future video coding technologies (after completing the Call for Evidence (CfE) in 2017).

The emerging video codec under the JVET development was titled “Versatile Video Coding”, or in short, VVC [42].

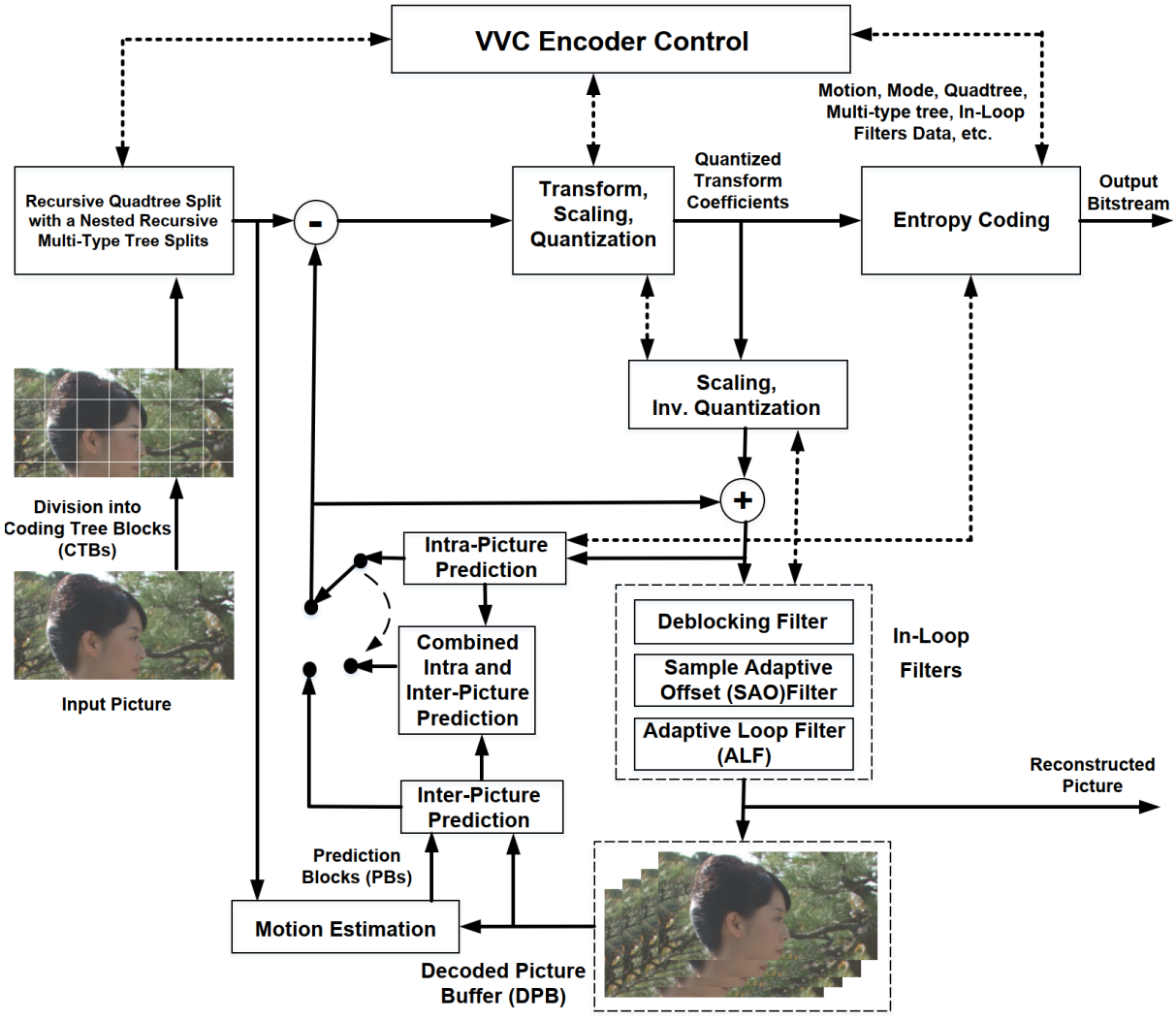

Figure 6 – The schematic block diagram of the H.266/MPEG-VVC encoder.

In turn, the first VVC draft along with the VVC Test Model 1 (VTM1) was published right after the April, 2018 meeting. As one of the main tools that provides a significant coding gain, VVC includes a quadtree with nested multi-type tree by using a coding block structure of binary and ternary splits. Further, additional tools and features include: Intra-mode coding with 67 Intra-picture prediction modes; Intra Block Copy (IBC); Bi-Directional Optical Flow (BDOF); Adaptive Motion Vector Resolution (AMVR); Geometric Partitioning Mode (GPM); Combined Inter and Intra Prediction (CIIP); Adaptive Loop Filter (ALF); *and many others*. The first version of the VVC standard (i.e., VVC v1) was officially finalized during the 19th JVET meeting, which took place between June 22 and July 1, 2020, and the VVC codec is currently starting to be widely deployed worldwide [42]. The schematic block diagram of the H.266/MPEG-VVC encoder is presented in *Figure 7*.

Therefore, the perceptual quantization matrices and masking encoding schemes described in this paper, can be applied during the VVC encoding process as well, thereby leading to even larger coding gains.

In addition, the presented perceptual video coding optimizations can be applied to the framework of [43], as described below. According to this framework, the x265-based multi-resolution encoding architecture is leveraged to significantly speed-up the encodes through sharing the analysis information from lower to higher resolutions. For example, the encodes that generate representations at the 960x540 resolution (i.e., 540p) can share information to those that generate representations at the 1920x1080 resolution (i.e., 1080p, a dyadic multiple of 540p), which in-turn can speed up the 3840x2160 resolution encode (i.e., 2160p, a dyadic multiple of 1080p). In addition, the encodes that generate representations for the 1280x720 resolution can share information to the encode that generates a representation at the 2560x1440 resolution.

Therefore, a framework can run efficient encodes, which generate both the reference and dependent representations in parallel, while also efficiently handling dependencies with the sharing analysis information among these presentations. In turn, the Adaptive Bit-Rate (ABR) ladder allows the more efficient sharing of the aforementioned analysis information among the representations at different resolutions.

Further, *Figure 7* schematically presents a high-level block-diagram of the proposed framework for the efficient usage of the x265-based [28] adaptive multi-resolution encoding architecture, thereby enabling the efficient sharing of analysis information across representations. As shown in *Figure 7*, the proposed framework is being used for converting an input video source, which has the WxH resolution, to n representations, each representation having a scaling factor of L_i *(with* $i = 0$ *to* n*)* and bitrate B_i. This framework uses a configuration file in order to define the encoding graph and the degree of encoder decision reuse between a dependent and a reference representation. In addition, it is possible to add representation-specific parameters to the configuration file, such as limiting the Coding Tree Unit (CTU) size for lower resolutions.

For the more detailed description regarding the syntax of the x265 configuration file, the reader is referred to [28].

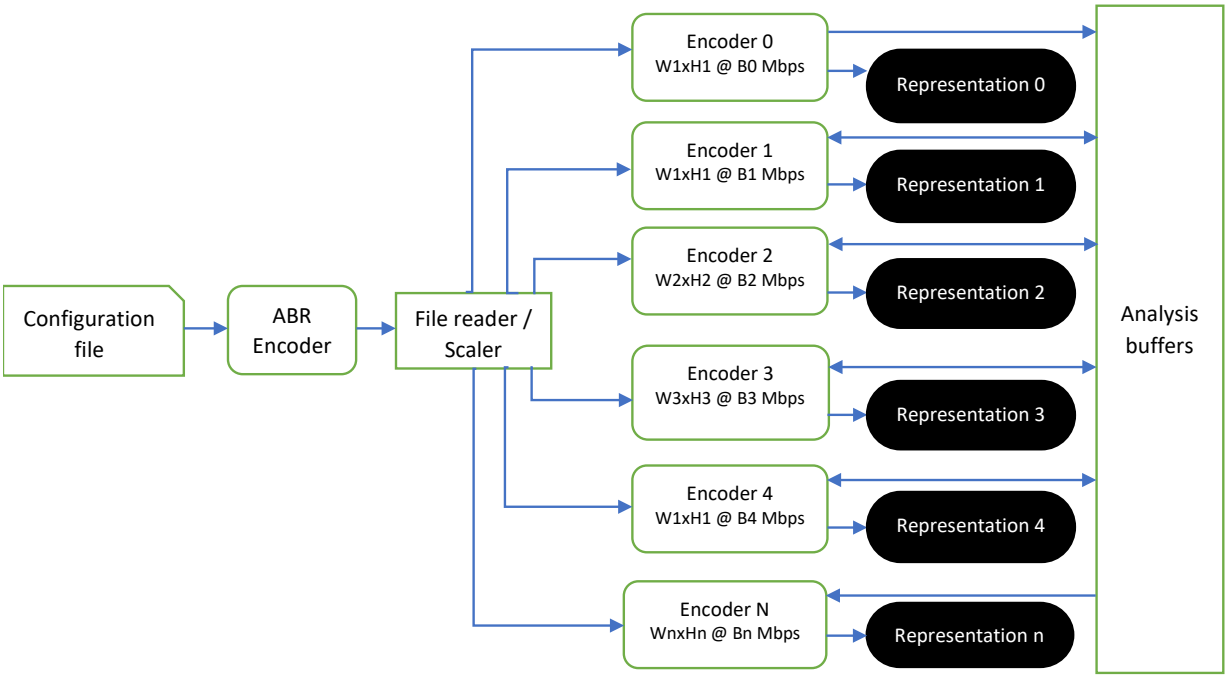

Figure 7 – A multi-resolution encoding framework for enabling efficient sharing of analysis information across representations.

By such a way, i.e. by utilizing perceptual quantization matrices and masking encoding schemes with VVC and with the above-mentioned multi-resolution encoding framework, further significant bit-rate savings can be achieved.

6. Conclusion

In this paper, perceptual video coding optimization techniques, including most recent trends and future directions, have been discussed in details. A special emphasis was made on the UltraHD resolution, such as the 3840x2160 (4K) resolution in terms of luma samples, and on the most recent, and currently the most advanced, H.265/MPEG-HEVC video coding standard. The development of the perceptual quantization matrices has been motivated by the Daly HVS-based perceptual model, which was further fitted into the more advanced and more complex Barten model (that incorporates a variety of HVS parameters) for much more accurate generation of these matrices. As a result, the video transmission bit-rate for UltraHD displays was reduced up to 11.3% in terms of SSIMPlus, while keeping the visual quality at substantially the same level. On the other hand, for smaller size mobile device displays, the video transmission bit-rate was reduced up to about 25% in terms of SSIMPlus objective quality metric, while keeping the visual quality substantially at the same level. In addition, a joint backward and forward temporal masking framework was presented, which considers temporal distances between frames and the closest scenecuts. This framework has been implemented in the popular x265 HEVC-based encoder. Based on the extensive subjective quality assessments, significant bitrate savings of up to about 26% are achieved for substantially the same perceived visual quality. The future direction mostly refer to implementing the presented framework with the most recent VVC video coding standard, further getting benefit from the presented x265-based multi-resolution encoding architecture.

Abbreviations

ABR	Adaptive Bit-Rate
ALF	Adaptive Loop Filter
AMVR	Adaptive Motion Vector Resolution
AVC	Advanced Video Coding
BDOF	Bi-Directional Optical Flow
CfE	Call for Evidence
CfP	Call for Proposals
CIIP	Combined Inter and Intra Prediction
CSF	Contrast Sensitivity Function
CTC	Common Test Conditions
CTU	Coding Tree Unit
DCT	Discrete Cosine Transform
DPCM	Differential Pulse Code Modulation
fps	frame per second
GOP	Group of Pictures
HDR	High Dynamic Range
HEVC	High Efficiency Video Coding
HVS	Human Visual System
IBC	Intra Block Copy
JCT-VC	Joint Collaborative Team on Video Coding
JND	Just Noticeable Difference
JVET	Joint Video Exploration Team
MPEG	Moving Pictures Experts Group
PPS	Picture Parameter Set
QM	Quantization Matrix
QP	Quantization Parameter
SDR	Standard Dynamic Range
SPS	Sequence Parameter Set
VCEG	Video Coding Experts Group
VTM	VVC Test Model
VVC	Versatile Video Coding

Bibliography & References

[1] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, "Performance comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders," *Picture Coding Symposium (PCS) 2013*, pp.394-397, 8-11 Dec. 2013.
[2] D. Grois, D. Marpe, T. Nguyen, and O. Hadar, "Comparative Assessment of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC Encoders for Low-Delay Video Applications", Proc. SPIE 9217, *Applications of Digital Image Processing XXXVII,* 92170Q, Sept. 2014.
[3] D. Grois, T. Nguyen, and D. Marpe, "Coding Efficiency Comparison of AV1/VP9, H.265/MPEG-HEVC, and H.264/MPEG-AVC Encoders," *Picture Coding Symposium (PCS)*, Dec. 2016.
[4] D. Grois, T. Nguyen, and D. Marpe,"Performance Comparison of AV1, JEM, VP9, and HEVC Encoders", Proc. SPIE 10396, *Applications of Digital Image Processing XL*, 103960L, 7-10 Aug., 2017.
[5] D. Grois *et al*., "Performance Comparison of Emerging EVC and VVC Video Coding Standards with HEVC and AV1," in *SMPTE Motion Imaging Journal*, vol. 130, no. 4, pp. 1-12, May 2021.
[6] ITU-T, Recommendation H.265 (04/13), Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services – Coding of Moving Video, High Efficiency Video Coding.
[7] ITU-T, Recommendation H.265 (10/14), Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services – Coding of Moving Video, High Efficiency Video Coding.
[8] ITU-T, Recommendation H.265 (04/15), Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services – Coding of Moving Video, High Efficiency Video Coding.
[9] ITU-T, Recommendation H.265 (12/16), Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services – Coding of Moving Video, High Efficiency Video Coding.
[10] D. Grois, and A. Giladi, "Perceptual quantization matrices for high dynamic range H.265/MPEG-HEVC video coding", Proc. SPIE 11137, Applications of Digital Image Processing XLII, 111370O, 2020.
[11] D. Grois, and A. Giladi, "HVS-Based Perceptual Quantization Matrices For HDR HRVC Video Coding for Mobile Devices", pp. 1-14, IBC 2020.
[12] D. Grois, A. Giladi, P. K. Karadugattu, and N. Balasubramanian, "Novel temporal masking framework for perceptually optimized video coding", *In Proceedings of the 1st Mile-High Video Conference (MHV '22)*. Association for Computing Machinery, New York, NY, USA, 119–120.
[13] G. M. Johnson and M. D. Fairchild, "On Contrast Sensitivity in an Image Difference Model", *In Proceedings of the IS&T PICS Conference*, pp. 18- 23, Portland, OR, 2002.
[14] B.A. Wandell, "Foundation of Vision," Sinear Associates, Sunderland, MA, 1995.
[15] S. Daly, "Subroutine for the generation of a two dimensional human visual contrast sensitivity function," Technical Report 233203Y, Eastman Kodak, 1987.
[16] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," In A. B. Watson, editor, *Digital Images and Human Vision*, pp. 179–206, 1993.
[17] S. Daly, "A visual model for optimizing the design of image processing algorithms," *International Conference on Image Processing (ICIP)*, vol. 2, pp. 16–20, Nov.1994.
[18] S. Daly, T. Kunkel, X. Sun, S. Farrell, and P. Crum, "Preference limits of the visual dynamic range for ultra high quality and aesthetic conveyance," *Proceedings of the SPIE*, 8651:86510J–86510J–11, 2013.
[19] J. L. Mannos, and D. J. Sakrison, "The Effects of a Visual Fidelity Criterion on the Encoding of Images," IEEE Trans. on Info. Theory, Vol. IT-20, No. 4, July 1974.
[20] R. Rosenholtz and A. B. Watson, "Perceptual adaptive JPEG coding," In Proceedings of the IEEE International Conference on Image Processing (ICIP), vol. 1, pp. 901–904, Sep. 1996.
[21] L.W. Chang, C.Y. Wang and S.M. Lee, "Designing JPEG quantization tables based on human visual system," *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, Kobe, 1999, pp. 376-380 vol.2.
[22] M. Nezamabadi, S. Miller, S. Daly, and R. Atkins, "Color signal encoding for high dynamic range and wide color gamut based on human perception," *Proceedings of the SPIE*, 9015:90150C–90150C–12, 2014.
[23] P. G. J. Barten. Physical model for the contrast sensitivity of the human eye. Proceedings of the SPIE, 1666:57–72, 1992.
[24] P. G. J. Barten. "Contrast sensitivity of the human eye and its effects on image quality," volume 72, *SPIE press*, USA, Dec. 1999.
[25] P. G. J. Barten, "Formula for the contrast sensitivity of the human eye," *Proceedings of the SPIE*, 5294:231–238, 2003.

[26] IMDB content database, Online: https://www.imdb.com/title/tt2872732.
[27] UltraHD Forum, Online: https://ultrahdforum.org.
[28] Projects from VideoLAN, x265 software library and application, Online: https://www.videolan.org/developers/x265.html.
[29] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves", ITU-T Q.6/SG16 VCEG 13th Meeting, Document VCEG-M33, Austin, USA, Apr. 2001.
[30] Eurofins®, "Which is the best objective video quality measure and why use SSIMPLUS", Eurofins®, Online: https://cdnmedia.eurofins.com/digitaltesting/media/116613/qoe-why-ssimplus.pdf .
[31] SSIMWAVE®, "SSIMPLUS Outperforms VMAF", 2017.
[32] F. Bossen, "Common HM test conditions and software reference configurations," document JCTVC-L1100 of JCT-VC, Geneva, CH, Jan. 2013.
[33] V. Adzic, H. S. Hock, and H. Kalva, "Visually lossless coding based on temporal masking in human vision," Proc. SPIE 9014, Human Vision and Electronic Imaging XIX, 90141C (25 February 2014)
[34] V. Adzic, H. Kalva and B. Furht, "Exploring visual temporal masking for video compression," *2013 IEEE International Conference on Consumer Electronics (ICCE)*, 2013, pp. 590-591.
[35] Projects from VideoLAN, x264 software library and application, Online: https://www.videolan.org/developers/x264.html.
[36] V. Adzic, H. Kalva and B. Furht, "Temporal visual masking for HEVC/H.265 perceptual optimization," *2013 Picture Coding Symposium (PCS)*, 2013, pp. 430-433.
[37] Siddique AA, Qadr MT, Mohy-Ud-Din Z. Masking of temporal activity for video quality control, measurement and assessment. Measurement and Control. 2020;53(9-10):1817-1824.
[38] K.-C. Liu, "Color Video JND Model Using Compound Spatial Masking and Structure-Based Temporal Masking," in *IEEE Access*, vol. 8, pp. 136760-136768, 2020.
[39] H. Wang, L. Yu, H. Yin, T. Li, and S. Wang, "An improved DCT-based JND estimation model considering multiple masking effects," Journal of Visual Communication and Image Representation, vol. 71, 102850, 2020.
[40] ITU-R, Recommendation ITU-R BT.500-14 (10/2019), Methodologies for the subjective assessment of the quality of television images.
[41] M. Wien, and V. Baroncini, "Status Report on SDR HD Verification Test Preparation", Doc. JVET-T0043, Teleconference, 7-16 Oct. 2020.
[42] ITU-T, Recommendation H.266 (08/2020), Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services – Coding of Moving Video, Versatile Video Coding.
[43] A. Mathesawaran, P. K. Karadugattu, P. Ramachandran, A. Giladi, D. Grois, P. Venkatesan, and A. Balk, "Open source framework for reduced-complexity multi-rate HEVC encoding," *Proc. SPIE 11510, Applications of Digital Image Processing XLIII*, 115101Y (25 August 2020);

