International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Series H

Supplement 15

(01/2017)

SERIES H: AUDIOVISUAL AND MULTIMEDIA SYSTEMS

Conversion and coding practices for HDR/ WCG Y'CbCr 4:2:0 video with PQ transfer characteristics

ITU-T H-series Recommendations – Supplement 15

ITU-T

International Telecommunication Union

ITU-T H-SERIES RECOMMENDATIONS

AUDIOVISUAL AND MULTIMEDIA SYSTEMS

For further details, please refer to the list of ITU-T Recommendations.

Supplement 15 to ITU-T H-series Recommendations

Conversion and coding practices for HDR/WCG Y′CbCr 4:2:0 video with PQ transfer characteristics

Summary

Supplement 15 to the ITU-T H-series of Recommendations provides guidance on the processing of high dynamic range (HDR) and wide colour gamut (WCG) video content. The purpose of this document is to provide a set of publicly referenceable recommended guidelines for the operation of advanced video coding (AVC) or high efficiency video coding (HEVC) video coding systems adapted for compressing HDR/WCG video for consumer distribution applications. This document includes a description of processing steps for converting from 4:4:4 RGB linear light representation video signals into non-constant luminance (NCL) Y′CbCr video signals that use the perceptual quantizer (PQ) transfer function defined in SMPTE ST 2084 and Recommendation ITU-R BT.2100. Although the focus of this document is primarily on 4:2:0 Y′CbCr 10 bit representations, these guidelines may also apply to other representations with higher bit depth or other colour formats, such as 4:4:4 Y′CbCr 12 bit video. In addition, this document provides some high-level recommendations for compressing these signals using either the AVC or HEVC video coding standards. A description of post-decoding processing steps is also included for converting these NCL Y′CbCr signals back to a linear light, 4:4:4 RGB representation.

This Supplement was jointly developed by ITU-T Study Group 16 and ISO/IEC JTC1 SC29/WG11 and the ISO title is ISO/IEC TR 23008-14 "*Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 14: Conversion and coding practices for HDR/WCG Y'CbCr 4:2:0 video with PQ transfer characteristics*".

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T H Suppl. 15	2017-01-27	16	11.1002/1000/13243

Keywords

AVC, HDR, HEVC, video coding, WCG.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this publication, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this publication is voluntary. However, the publication may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the publication is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the publication is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this publication may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the publication development process.

As of the date of approval of this publication, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this publication. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Introduction

High dynamic range (HDR) video is a type of video content in which the sample values span a larger luminance range than conventional standard dynamic range (SDR) video. HDR video can provide an enhanced viewer experience and can more accurately reproduce scenes that include, within the same image, dark areas and bright highlights, such as emissive light sources and reflections. Wide colour gamut (WCG) video, on the other hand, is video characterized by a wider spectrum of colours compared to what has been commonly available in conventional video. Recent advances in capture and display technology have enabled consumer distribution of HDR and WCG content. However, given the characteristics of such content, special considerations may need to be made, in terms of both processing and compression, compared to conventional content.

This Supplement provides a set of recommended guidelines on the processing of consumer distribution HDR/WCG video. This includes recommendations for converting a video signal, in a linear light RGB representation with ITU-R BT.2020 colour primaries, to a 10-bit, narrow range, PQ encoded (as defined in SMPTE ST 2084 and Recommendation ITU-R BT.2100), 4:2:0, non-constant luminance Y′CbCr representation. These guidelines may also apply to other representations with a higher bit depth or other colour formats, such as 4:4:4 Y′CbCr 12 bit video. The scope of this document is illustrated in Figure 1.

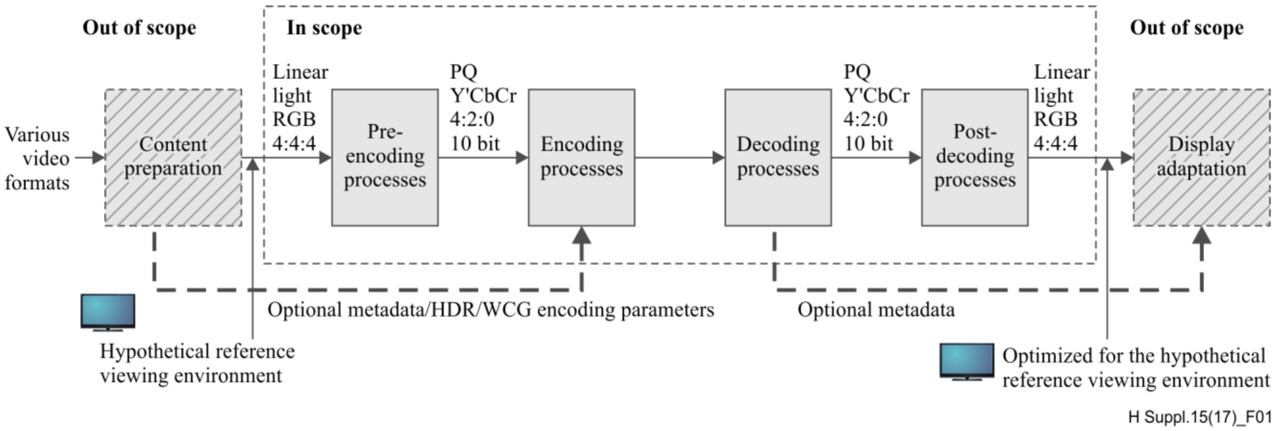

Figure 1 – Illustration of the scope of this document

The content preparation step, as well as the display adaptation step, are considered to be out of the scope of this document. However, metadata generated during the content preparation step may be passed through the encoder-decoder chain and can significantly affect the display adaptation step. The content preparation step may include filtering and image enhancement processing such as de-noising, colour correction and sharpening, as well as other processes. Such methods are deliberately not described in this document. The processing steps described in this document are made available for reference only and the document does not contain any elements of normative nature. It is possible to replace one or more of the processing steps described in this document, for example, in order to reduce computational complexity or to improve fidelity. This document's intention is to provide some guidelines for operating an HDR/WCG video system that is constrained to code a 10-bit, PQ (as defined in [SMPTE ST 2084] and [ITU-R BT.2100-0], 4:2:0, non-constant luminance Y′CbCr signal representation. This configuration is also aligned with the HDR10 media profile defined in [DECE], the interface defined in [CTA 861-G] and the restrictions in [Blu-ray2015]. The processing steps in this document are optimized with the intention of providing the best possible result when the same hypothetical reference viewing environment (HRVE) is used before and after the HDR/WCG system. This document does not account for when the viewing environment used after the HDR/WCG system is different from the viewing environment used as the HRVE. In particular, display adaptation, such as the techniques described in the SMPTE ST 2094 standards, is not considered in this document. [ITU-R BT.2390-0] contains additional information on viewing environments and examples of parameters that may be appropriate to apply for practical HDR/WCG systems. This document does not provide a description of any preferred HRVE, but acknowledges the fact that in many applications of HDR/WCG video it may be desirable to have a well-defined HRVE description in order to ensure alignment between content preparation and content consumption.

Supplement 15 to the H-Series of Recommendations

Conversion and coding practices for HDR/WCG Y′CbCr 4:2:0 video with PQ transfer characteristics

1 Scope

This Supplement provides guidance on the processing of high dynamic range (HDR) and wide colour gamut (WCG) video content. The purpose of this document is to provide a set of publicly referenceable recommended guidelines for the operation of AVC or HEVC video coding systems adapted for compressing HDR/WCG video for consumer distribution applications. This document includes a description of processing steps for converting from 4:4:4 RGB linear light representation video signals into non-constant luminance (NCL) Y′CbCr video signals that use the perceptual quantizer (PQ) transfer function defined in [SMPTE ST 2084] and [ITU-R BT.2100-0]. Although the focus of this document is primarily on 4:2:0 Y′CbCr 10 bit representations, these guidelines may also apply to other representations with higher bit depth or other colour formats, such as 4:4:4 Y′CbCr 12 bit video. In addition, this document provides some high-level recommendations for compressing these signals using either the AVC or HEVC video coding standards. A description of post-decoding processing steps is also included for converting these NCL Y′CbCr signals back to a linear light, 4:4:4 RGB representation.

2 References

[ITU-T H.264] Recommendation ITU-T H.264 (in force) | ISO/IEC 14496-10 (in force), *Advanced video coding for generic audiovisual services*.

[ITU-T H.265] Recommendation ITU-T H.265 (in force) | ISO/IEC 23008-2 (in force)*High Efficiency Video Coding.*

[ITU-R BT.709-6] Recommendation ITU-R BT.709-6 (2015), *Parameter values for the HDTV standards for production and international programme exchange*.

[ITU-R BT.1886-0] Recommendation ITU-R BT.1886-0 (2011), *Reference electro-optical transfer function for flat panel displays used in HDTV studio production*.

[ITU-R BT.2020-2] Recommendation ITU-R BT.2020-2 (2015), *Parameter values for ultra-high definition television systems for production and international programme exchange*.

[ITU-R BT.2100-0] Recommendation ITU-R BT.2100-0 (2016), *Image parameter values for high dynamic range television for use in production and international programme exchange*.

[ITU-R BT.2390-0] Report ITU-R BT.2390-0 (2016), *High dynamic range television for production and international programme exchange*.

[Baroncini 2016] V. Baroncini, K. Andersson, A. K. Ramasubramonian, and G. J. Sullivan (editors) (2016), *Revised Verification Test Report for HDR/WCG Video Coding Using HEVC Main 10 Profile*, JCTVC-Y1018.

[Blu-ray 2015] Blu-ray Disc Association (2015), *BD-ROM: Audio Visual Application Format Specifications version 3*
[CTA 861-G] CTA 861-G (2017), *A DTV Profile for Uncompressed High Speed Digital Interfaces*.

[DECE] DECE (2015), *Common File Format & Media Formats Specification Version 2.1.*
<http://www.uvcentral.com/sites/default/files/files/PublicSpecs/CFFMediaFormat-2_1.pdf>
[Luthra 2015] A. Luthra, E. Francois, and W. Husak (editors) (2015), "Call for Evidence (CfE) for HDR and WCG Video Coding", ISO/IEC JTC 1/SC 29/WG 11 (MPEG) document N15083, Geneva, 2015

[Norkin 2016] A. Norkin (2016), *Fast algorithm for HDR video pre-processing*, in Proceedings of the IEEE Picture Coding Symposium (PCS), Nuremberg.

[Poynton 1996] C. Poynton (1996), *A Technical Introduction to Digital Video*, New York: John Wiley & Sons, ISBN-13: 978-0471122531.

[SMPTE RP 431-2] SMPTE RP 431-2 (2011), *D-cinema Quality – Reference Projector and Environment*.

[SMPTE ST 2084] SMPTE ST 2084 (2014), *High Dynamic Range Electro-Optical Transfer Function for Mastering Reference Display*.

[SMPTE ST 2094] SMPTE ST 2094 (2016), *Dynamic Metadata for Color Transforms of HDR and WCG Images*.

[Strom DCC2016] J. Ström, J. Samuelsson and K. Dovstam (2016), *Luma Adjustment for High Dynamic Range Video*, in Proceedings of the IEEE Data Compression Conference (DCC), Snowbird.

[Strom PCS2016] J. Ström, K. Andersson, M. Pettersson, P. Hermansson, J. Samuelsson, A. Segall, J. Zhao, S-H. Kim, K. Misra, A. M. Tourapis, Y. Su and D. Singer (2016), *High Quality HDR Video Compression using HEVC Main 10 Profile,* in Proceedings of the IEEE Picture Coding Symposium (PCS), Nuremberg.

3 Definitions

This Supplement defines the following terms. The definitions used in the AVC [ITU-T H.264] and HEVC [ITU-T H.265] standards also apply.

3.1 electro-optical transfer function (EOTF): The function used in the post-decoding process to convert from a non-linear representation to a linear representation.

3.2 full range: A range in a fixed-point (integer) representation that spans the full range of values that could be expressed with that bit depth, so that for 10-bit signals, black corresponds to code value 0 and peak white corresponds to code value 1023 for Y′, as per the full range definition from [ITU-R BT.2100-0].

3.3 inverse electro-optical transfer function (inverse EOTF): A function used in the pre-encoding process to convert from a linear representation to a non-linear representation, computed as the inverse of the EOTF.

NOTE – In this document the pre-encoding process is assumed to operate on HDR/WCG video content that has been prepared for a hypothetical reference viewing environment as shown in Figure 1. The content preparation step may contain processing such as applying an opto-optical transfer function (OOTF), in which the HDR/WCG video is converted from one linear representation (corresponding to the scene) to another linear representation (corresponding to the display). The OOTF has the role of applying a "rendering intent". In systems where no such OOTF is applied in the content preparation step, the process of converting from a linear representation (corresponding to the scene) to a non-linear representation is typically called the opto-electrical transfer function (OETF).

3.4 narrow range: A range in a fixed-point (integer) representation that does not span the full range of values that could be expressed with that bit depth, so that for 10 bit representations, the range from 64 (black) to 940 (peak white) is used for Y′, and the range from 64 to 960 is used for Cb and Cr, as per the narrow range definition from [ITU-R BT.2100-0].

NOTE – Narrow range is, in some applications, called by synonyms such as: "limited range", "video range", "legal range", "SMPTE range" or "standard range".

3.5 opto-electrical transfer function (OETF): The function that converts linear scene light into the video signal, typically within a camera.

3.6 opto-optical transfer function (OOTF): A function that maps relative scene linear light (typically the camera output signal) to display linear light (typically, the signal driving a mastering monitor).

3.7 random access point access unit (RAPAU): An access unit in the bitstream containing an intra-coded picture with the property that all pictures following the intra-coded picture in output order can be correctly decoded without using any information preceding the random access point access unit in the bitstream.

3.8 transfer function: In this document, a transfer function refers to any of the following; EOTF, inverse EOTF, OETF, inverse OETF, OOTF or inverse OOTF.

4 Abbreviations and acronyms

This Supplement uses the following abbreviations and acronyms:

AVC Advanced Video Coding [ITU-T H.264]

CL Constant Luminance

EOTF Electro-Optical Transfer Function

FIR Finite Impulse Response

HD High Definition

HDR	High Dynamic Range
HEVC	High Efficiency Video Coding [ITU-T H.265]
HRVE	Hypothetical Reference Viewing Environment
HVS	Human Visual System
IQU	Up-scaling and Inverse Quantization
LUT	Look-Up Table
MAD	Mean Absolute Difference
NCL	Non-Constant Luminance
PQ	Perceptual Quantizer (as defined in [SMPTE ST 2084] and [ITU-R BT.2100-0]
QD	Quantization, Down-scaling
QP	Quantization Parameter
RAPAU	Random Access Point Access Unit
RGB	colour system using Red, Green and Blue components
SSE	Sum of Squared Errors
SDR	Standard Dynamic Range
SEI	Supplemental Enhancement Information
SPS	Sequence Parameter Set
OETF	Opto-Electrical Transfer Function
OOTF	Opto-Optical Transfer Function
VUI	Video Usability Information
WCG	Wide Colour Gamut
XYZ	The CIE 1931 colour space. Y corresponds to the luminance signal.
Y′CbCr	Colour space representation commonly used for video/image distribution as a way of encoding RGB information, also commonly expressed as YCbCr, $Y'C_BC_R$, or $Y'C'_BC'_R$. The relationship between Y′CbCr and RGB is dictated by certain signal parameters, such as colour primaries, transfer characteristics and matrix coefficients. Unlike the (constant luminance) Y component in the XYZ representation, Y′ in this representation might not be representing the same quantity. Y′ is commonly referred to as "luma". Cb and Cr are commonly referred to as "chroma".

5 Conventions

5.1 General

The mathematical operators used in this document are similar to those used in the C programming language. However, the results of integer division and arithmetic shift operations are defined more precisely and additional operations are defined, such as exponentiation and real-valued division. Numbering and counting conventions generally begin from 0, e.g., "the first" is equivalent to the 0-th, "the second" is equivalent to the 1-th, etc.

5.2 Arithmetic operators

The following arithmetic operators are defined as follows:

+	Addition
−	Subtraction (as a two-argument operator) or negation (as a unary prefix operator)
*	Multiplication, including matrix multiplication
x^y	Exponentiation. Denotes x to the power of y. In other contexts, such notation is used for superscripting not intended for interpretation as exponentiation.

/ Integer division with truncation of the result towards zero. For example, 7 / 4 and (−7) / (−4) are truncated to 1 and (−7) / 4 and 7 / (−4) are truncated to −1.

÷ Used to denote division in mathematical formulae where no truncation or rounding is intended.

$\frac{x}{y}$ Used to denote division in mathematical formulae where no truncation or rounding is intended.

$\sum_{i=x}^{y} f(i)$ The summation of f(i) with i taking all integer values from x up to and including y.

x % y Modulus. Remainder of x divided by y, defined only for integers x and y with x >= 0 and y > 0.

5.3 Bit-wise operators

The following bit-wise operators are defined as follows:

& Bit-wise "and". When operating on integer arguments, operates on a two's complement representation of the integer value. When operating on a binary argument that contains fewer bits than another argument, the shorter argument is extended by adding more significant bits equal to 0.

| Bit-wise "or". When operating on integer arguments, operates on a two's complement representation of the integer value. When operating on a binary argument that contains fewer bits than another argument, the shorter argument is extended by adding more significant bits equal to 0.

^ Bit-wise "exclusive or". When operating on integer arguments, operates on a two's complement representation of the integer value. When operating on a binary argument that contains fewer bits than another argument, the shorter argument is extended by adding more significant bits equal to 0.

x >> y Arithmetic right shift of a two's complement integer representation of x by y binary digits. This function is defined only for non-negative integer values of y. Bits shifted into the MSBs as a result of the right shift have a value equal to the MSB of x prior to the shift operation.

x << y Arithmetic left shift of a two's complement integer representation of x by y binary digits. This function is defined only for non-negative integer values of y. Bits shifted into the LSBs as a result of the left shift have a value equal to 0.

5.4 Assignment operators

The following assignment operators are defined as follows:

= Assignment operator

++ Increment, i.e., x++ is equivalent to x = x + 1; when used in an array index, evaluates to the value of the variable prior to the increment operation.

− − Decrement, i.e., x− − is equivalent to x = x − 1; when used in an array index, evaluates to the value of the variable prior to the decrement operation.

+= Increment by amount given, i.e., x += 3 is equivalent to x = x + 3, and x += (−3) is equivalent to x = x + (−3).

−= Decrement by amount given, i.e., x −= 3 is equivalent to x = x − 3, and x −= (−3) is equivalent to x = x − (−3).

5.5 Relational, logical and other operators

The following operators are defined as follows:

== Equality operator

!= Not equal to operator

!x Logical negation "not"

> Larger than operator

< Smaller than operator

≥ Larger than or equal to operator

≤ Smaller than or equal to operator

&& Conditional/logical "and" operator. Performs a logical "and" of its Boolean operators, but only evaluates the second operand if necessary.

||	Conditional/logical "or" operator. Performs a logical "or" of its Boolean operators, but only evaluates the second operand if necessary.
a ? b : c	Ternary conditional. If condition a is true, then the result is equal to b; otherwise the result is equal to c.

5.6 Mathematical functions

The following mathematical functions are defined as follows:

$$\mathrm{Abs}(x) = \begin{cases} x & ; \quad x \geq 0 \\ -x & ; \quad x < 0 \end{cases}$$

$\mathrm{Ceil}(x)$ the smallest integer greater than or equal to x.

$$\mathrm{Clip3}(x, y, z) = \begin{cases} x & ; \quad z < x \\ y & ; \quad z > y \\ z & ; \quad \text{otherwise} \end{cases}$$

$\mathrm{Floor}(x)$ the largest integer less than or equal to x.

$\mathrm{EOTF}^{-1}(x)$ the inverse EOTF used to convert a linear light representation to a non-linear light representation.

$$\mathrm{Max}(x, y) = \begin{cases} x & ; \quad x > y \\ y & ; \quad \text{otherwise} \end{cases}$$

$$\mathrm{Max}(x, y, z) = \begin{cases} x & ; \quad x > \mathrm{Max}(y, z) \\ y & ; \quad y > \mathrm{Max}(x, z) \\ z & ; \quad \text{otherwise} \end{cases}$$

$$\mathrm{Min}(x, y) = \begin{cases} x & ; \quad x < y \\ y & ; \quad \text{otherwise} \end{cases}$$

$$\mathrm{Min}(x, y, z) = \begin{cases} x & ; \quad x < \mathrm{Min}(y, z) \\ y & ; \quad y < \mathrm{Min}(x, z) \\ z & ; \quad \text{otherwise} \end{cases}$$

$$\mathrm{Round}(x) = \mathrm{Sign}(x) * \mathrm{Floor}(\mathrm{Abs}(x) + 0.5)$$

$$\mathrm{Sign}(x) = \begin{cases} 1 & ; \quad x > 0 \\ 0 & ; \quad x = 0 \\ -1 & ; \quad x < 0 \end{cases}$$

$\mathrm{EOTF}(x)$ the EOTF used to convert a non-linear light representation x to a linear light representation.

5.7 Order of operations

When order of precedence in an expression is not indicated explicitly by use of parentheses, the following rules apply:

– Operations of a higher precedence are evaluated before any operation of a lower precedence.

– Operations of the same precedence are evaluated sequentially from left to right.

Table 1 specifies the precedence of operations from highest to lowest; a higher position in the table indicates a higher precedence.

NOTE – For those operators that are also used in the C programming language, the order of precedence used in this document is the same as that used in the C programming language.

Table 1 – Operation precedence from highest (at top of table) to lowest (at bottom of table)

Operations (with operands x, y and z)
"x++", "x− −"
"!x", "−x" (as a unary prefix operator)
"x^y"
"x * y", "x / y", "x ÷ y", "$\frac{x}{y}$", "x % y"
"x + y", "x − y" (as a two-argument operator), "$\sum_{i=x}^{y} f(i)$"
"x << y", "x >> y"
"x < y", "x <= y", "x > y", "x >= y"
"x = = y", "x != y"
"x & y"
"x	y"
"x && y"
"x || y"
"x ? y : z"
"x..y"
"x = y", "x += y", "x −= y"

6 Overview

The HDR/WCG system described in this document consists of four major stages:

– a pre-encoding stage consisting of several preprocessing processes (clause 7),

– an encoding stage (clause 8),

– a decoding stage (clause 9), and

– a post-decoding stage, also consisting of several post-processing processes (clause 10).

These four stages are applied sequentially, with the output of one stage being used as input to the next stage according to the above-mentioned order.

It is assumed that both the input to and the output of the HDR/WCG system are 4:4:4, linear light, floating-point signals, in an RGB colour representation using the same colour primaries. The output signal is targeted to resemble the input video signal as closely as possible. Other video formats can be used as input to the HDR/WCG system by first converting them to the above-defined input signal representation. The HDR/WCG system described in this document is, in practice, a system for both HDR and WCG video since it is assumed that the input video is represented with colour primaries in accordance with [ITU-R BT.2020] and [ITU-R BT.2100].

Two different models, *the simple reference model* and *the enhanced reference model*, are described in this document for the pre-encoding and encoding processes. The *simple reference model* corresponds to the reference configuration used in the MPEG call for evidence (CfE) on HDR and WCG [Luthra 2015], while the *enhanced reference model* corresponds to a new reference configuration that was developed in MPEG following the CfE. Both of these models were tested in the JCT-VC verification test on HDR/WCG video coding using HEVC Main 10 Profile [Baroncini 2016]. For the decoding process and post-decoding processes, a single model is described.

The primary purpose of the pre-encoding process is to convert the video input from its 4:4:4 RGB linear light, floating-point signal representation to a signal that is suitable for a video encoder. The conversion to a non-linear representation is performed in an attempt to exploit the characteristics of the human visual system (HVS) that could allow the re-quantization of the signal at a limited precision.

NOTE 1 – For a fixed-point linear HDR/WCG video representation, approximately a 28-bit integer representation would be required to avoid introducing visible quantization/banding errors due to the 28 f-stop linear light dynamic range (0.000 05 cd/m^2 to 10 000 cd/m^2) that is spanned by the PQ EOTF. In practice, the input to the HDR/WCG system will typically be in a non-linear representation that would either need to be first converted to linear light data or be directly converted to a non-linear representation using the PQ EOTF.

It is assumed that encoding and decoding is performed in a 4:2:0, 10-bit representation. An encoder is expected to make the best use of the encoding tools available according to a particular specification, profile and level, given also the characteristics of the content and the limitations of the intended application and implementation. In particular, different encoding algorithms, such as algorithms for motion estimation, mode decision, rate allocation, rate control and post-filtering control among other aspects, may have to be considered when encoding HDR/WCG material, in a given representation, compared to SDR material. The decoding process, on the other hand, is fully specified in the respective HEVC and AVC video coding standards, and a decoder must fully comply with the intended profile and level to properly decode and output the reconstructed video samples from a given input bitstream.

NOTE 2 – The focus of this document is on consumer and direct-to-home applications, which are expected to use, at least in the near term future, a 4:2:0 10-bit format. Processes similar to the ones described in this document can be used for conversion and compression of other formats, such as 4:2:2 and 4:4:4 chroma formats or video with a bit depth higher than 10 bits.

The steps in the post-decoding process are aligned with what is commonly referred to as the non-constant luminance representation (NCL) in which colour conversion, to R′G′B′, is performed prior to applying the EOTF to produce linear RGB sample values.

There is no specific or minimum bit depth required for performing the operations described in the pre-encoding process and the post-encoding process. Using the precision associated with 64 bit floating-point operations will give high accuracy, but it is also possible to use fixed-point arithmetic or floating-point operations with precision lower than 64 bits. It is recommended to avoid using too low precision since it could potentially lead to a loss of precision in the output video. The input to the encoding step and the output of the decoding step are, however, 10-bit integer representations.

7 Pre-encoding process

7.1 General

The pre-encoding process described in this document includes the following components:

a) a conversion component from a linear data representation to a non-linear data representation using the appropriate EOTF,

b) a colour format conversion component that converts data to the non-constant luminance Y′CbCr representation,

c) a conversion component that converts a floating-point to a fixed-point representation (e.g., 10 bits), and

d) a chroma down-conversion component that converts data from 4:4:4 to 4:2:0.

NOTE – Picture resolution scaling may also be a vital component of the pre-encoding process; for example, if the target system requires a particular image resolution to be delivered to the decoder. It may be desirable, for example, to rescale a source from a 1920×1080 to a 3840×2160 resolution, or vice versa. Such scaling is not included in the scope of this document. However, it is common practice that, for improved performance even for SDR material, rescaling would be performed in the linear domain.

Figure 2 presents a diagram of how these components are combined in the *simple reference* model to generate the desirable outcome, in a conventional manner. In this model, all blocks work independently, whereas chroma subsampling is performed using fixed-point arithmetic and at the same precision as the target outcome.

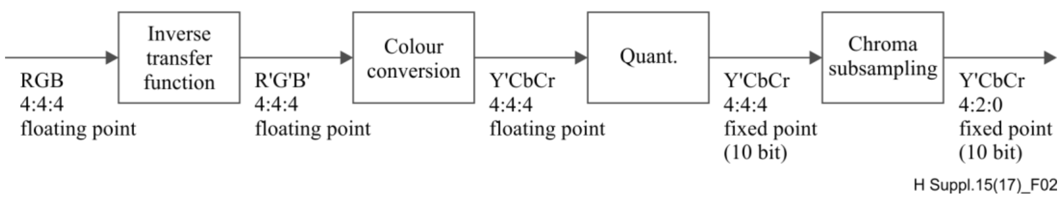

Figure 2 – Conventional pre-encoding process system diagram

Although this combination could be the most appropriate for some implementations, it has several limitations that can affect both its performance and implementation complexity. In this clause, the pre-encoding process components are first introduced in more detail, and then the alternative configuration corresponding to the *enhanced reference model* is presented in clause 7.3. Recommendations on how to best utilize some of the conversion components are also presented.

7.2 Pre-encoding process stages

7.2.1 Conversion from a linear to a non-linear light representation: RGB to R′G′B′

Conversion from a linear to a non-linear light representation is performed using an inverse EOTF, or as is commonly referred to in other specifications, an opto-electrical transfer function (OETF). In this document, the PQ EOTF defined in [SMPTE ST 2084] and [ITU-R BT.2100] is used.

More specifically, the non-linear light representation V of a linear light intensity signal L_o, that takes values normalized to the range [0, 1], can be computed as:

$$V = EOTF^{-1}(L_o) = \left(\frac{c_1 + c_2 * L_o^n}{1 + c_3 * L_o^n}\right)^m \quad (7\text{-}1)$$

where c_1, c_2, c_3, m and n are constants, which are defined as follows:

$$c_1 = c_3 - c_2 + 1 = 3\,424 \div 4\,096 = 0.835\,937\,5 \quad (7\text{-}2)$$

$$c_2 = 2\,413 \div 128 = 18.851\,562\,5 \quad (7\text{-}3)$$

$$c_3 = 299 \div 16 = 18.687\,5 \quad (7\text{-}4)$$

$$m = 2\,523 \div 32 = 78.843\,75 \quad (7\text{-}5)$$

$$n = 1\,305 \div 8\,192 = 0.159\,301\,757\,812\,5 \quad (7\text{-}6)$$

The peak value of 1 for L_o is ordinarily intended to correspond to an intensity level of 10 000 candelas per square metre (cd/m^2), while the value of 0 for L_o is ordinarily intended to correspond to an intensity level of 0 cd/m^2. The behaviour of the inverse PQ EOTF in relationship to the ITU-R BT.709 OETF and the inverse of the ITU-R BT.1886 EOTF is shown in Figure 3.

NOTE – A direct comparison of the inverse of the PQ EOTF with the ITU-R BT.709 OETF might not be appropriate since [ITU-R BT.709] may assume the use of an OOTF during decoding.

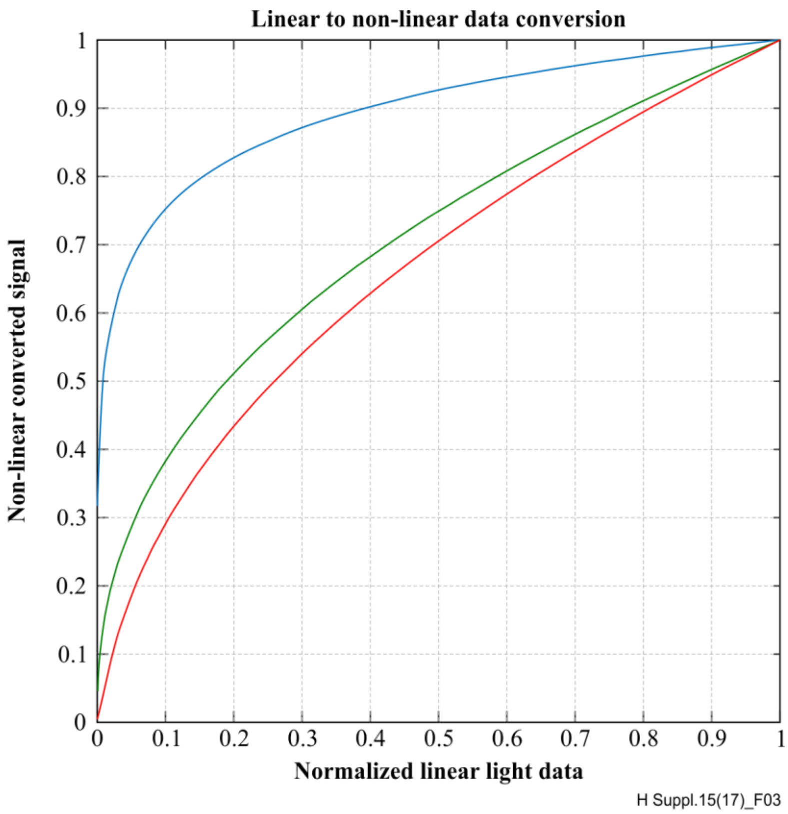

Figure 3 – Graph of the inverse of the PQ EOTF, the inverse of the ITU-R BT.1886 EOTF and the ITU-R BT.709 OETF

This process is applied to all R, G and B linear light samples, where each component is a number between 0.0 (representing no light) and 1.0 (representing 10 000 cd/m^2). This results in their non-linear counterparts R′, G′ and B′ as follows.

$$R' = \mathrm{EOTF}^{-1}(R) \qquad (7\text{-}7)$$

$$G' = \mathrm{EOTF}^{-1}(G) \qquad (7\text{-}8)$$

$$B' = \mathrm{EOTF}^{-1}(B) \qquad (7\text{-}9)$$

The resulting values for R′, G′ and B′ are numbers between 0.0 and 1.0. Although it is, in general, recommended to perform this conversion process using Formula 7-1 directly, this, however, may not be possible in some implementations given the complexity of the computation. Instead, look-up tables (LUT) may be preferred. Due to the characteristics of the conversion and the desire to achieve high precision for both low/dark and high values, it is highly recommended that, in such scenarios, a non-uniformly indexed LUT interpolator is used as described in [Strom PCS2016]. Such schemes can achieve relatively high accuracy/minimum approximation error for the conversion, while achieving considerable memory savings.

7.2.2 Colour representation conversion: R′G′B′ to non-constant luminance Y′CbCr

Conversion from the R′G′B′ to the non-constant luminance Y′CbCr representation is commonly performed using a 3×3 matrix conversion process of the form:

$$\begin{bmatrix} Y' \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} w_{YR} & w_{YG} & w_{YB} \\ w_{CbR} & w_{CbG} & w_{CbB} \\ w_{CrR} & w_{CrG} & w_{CrB} \end{bmatrix} \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \mathbf{W} * \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} \qquad (7\text{-}10)$$

where w_{YR}, w_{YG}, w_{YB}, w_{CbR}, w_{CbG}, w_{CbB}, w_{CbR}, w_{CbG} and w_{CbB} are constants. The values for w_{YR}, w_{YG} and w_{YB}, are set to exactly the same values used to convert R, G and B data to the CIE 1931 Y (luminance) signal. For ITU-R BT.2100 colour primaries, these are defined as follows:

$$w_{YR} = 0.262\ 7 \quad (7\text{-}11)$$

$$w_{YG} = 0.678\ 0 \quad (7\text{-}12)$$

$$w_{YB} = 0.059\ 3 \quad (7\text{-}13)$$

The resulting value for Y′ will be between 0.0 and 1.0. The values of the constants w_{CbR}, w_{CbG}, w_{CbB}, w_{CrR}, w_{CrG} and w_{CrB} are computed in a manner that the resulting Cb and Cr components are always within the [−0.5, 0.5] range. This results in the following values:

$$w_{CbR} = -\frac{w_{YR}}{2*(1-w_{YB})} = -0.139\ 630\ 063 \quad (7\text{-}14)$$

$$w_{CbG} = -\frac{w_{YG}}{2*(1-w_{YB})} = -0.360\ 369\ 937 \quad (7\text{-}15)$$

$$w_{CbB} = 0.5 \quad (7\text{-}16)$$

$$w_{CrR} = 0.5 \quad (7\text{-}17)$$

$$w_{CrG} = -\frac{w_{YG}}{2*(1-w_{YR})} = -0.459\ 785\ 705 \quad (7\text{-}18)$$

$$w_{CrB} = -\frac{w_{YB}}{2*(1-w_{YR})} = -0.040\ 214\ 295 \quad (7\text{-}19)$$

An alternative method to perform the same conversion process is presented in [ITU-R BT.2020] and [ITU-R BT.2100], where the chroma components are computed after the conversion of the luma component according to Formula 7-10 as follows:

$$\mathrm{Cb} = \frac{\mathrm{B}'-\mathrm{Y}'}{alpha} \quad (7\text{-}20)$$

$$\mathrm{Cr} = \frac{\mathrm{R}'-\mathrm{Y}'}{beta} \quad (7\text{-}21)$$

with $alpha = 2*(1-w_{YB})$ and $beta = 2*(1-w_{YR})$.

This can be seen as equivalent to the matrix presented in Formula 7-10.

The inverse process, i.e., converting Y′, Cb and Cr data back to R′, G′ and B′ data, is described in clause 10.4.

7.2.3 Chroma down-conversion

Converting the HDR/WCG video data from a 4:4:4 representation to a 4:2:0 representation nominally involves filtering and down-converting/subsampling the two chroma planes in both the horizontal and vertical directions. It is, though, possible to apply more complex chroma down-conversion methods that preserve edges and thus reduce the impact of interpolated colour values that did not exist in the local neighbourhood of a pixel in the original 4:4:4 representation. It is also a requirement, according to both [ITU-R BT.2020] and [ITU-R BT.2100], that the resulting chroma samples are co-sited with those of luma at even horizontal and vertical positions (Figure 4) where the first sample and line are counted starting from zero.

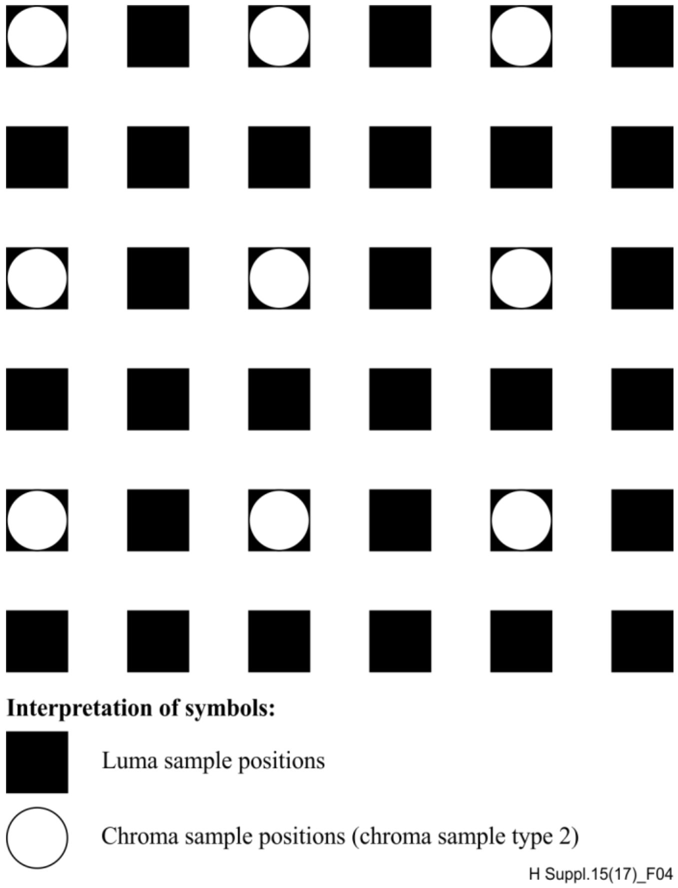

Figure 4 – Chroma and luma sample location relationship

It is anticipated that a considerable amount of consumer electronics conversion systems would use 2-D separable finite impulse response (FIR) linear filters for low-pass filtering the chroma data before subsampling (2:1 decimation step). Such filters would basically be of the form:

$$y[n] = \sum_{i=-N}^{N} b_i * x[n+i] \qquad (7\text{-}22)$$

where x[n] is the input chroma signal, y[n] is the filtered output chroma signal, (2 * N) corresponds to the filter order or, equivalently, (2 * N + 1) corresponds to the number of taps of the filter, and b_i corresponds to the coefficient of the filter at position i.

It has been observed that, especially due to the nonlinear characteristics of the PQ EOTF and its effect on quantization, special caution needs to be exercised when selecting the filter coefficients of such a resampling filter, in order to mitigate chroma "leakage" as defined in [Poynton 1996]. Conventional filters, such as linear filters, that are commonly used for down-conversion of SDR chroma signals may potentially result in visual artefacts when applied to HDR/WCG signals. This document however only considers two short-tap-length linear FIR filters, which have been used in experiments for the development of this report. Such filters can be utilized for both vertical and horizontal filtering of the chroma samples. Both *simple reference* and *enhanced reference* models use filter f_0:

Table 2 – Suggested filters for chroma down-sampling

Filter	Filter coefficients		
	b_{-1}	b_0	b_1
f_0	$\frac{1}{8}$	$\frac{6}{8}$	$\frac{1}{8}$
f_1	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

The characteristics (magnitude and phase) of these filters are shown in Figure 5. Filter f_1 has a stronger attenuation that is equal to −6dB at 0.5π rad/s, whereas filter f_0 could potentially cause some aliasing artefacts due to having a significant amount of energy remaining in its stop-band.

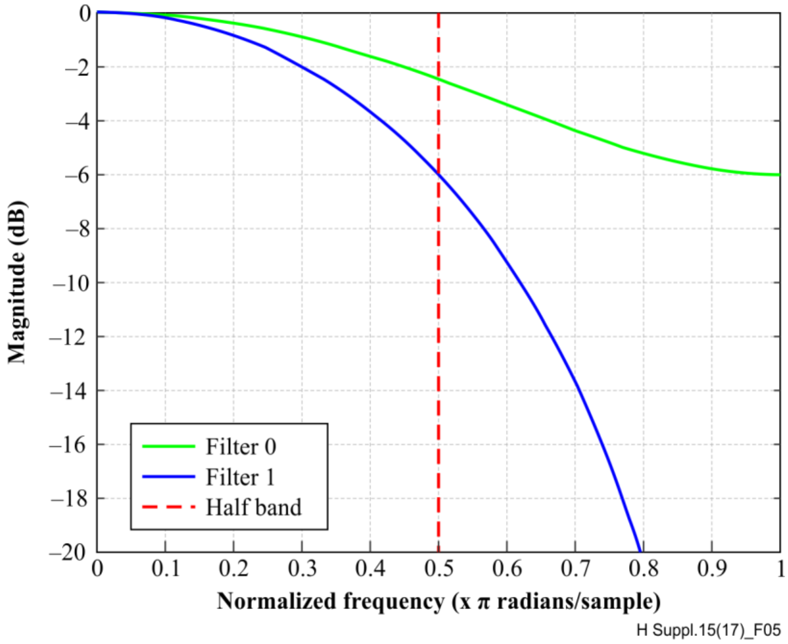

Figure 5 – Frequency response of filters f_0 and f_1 for chroma down-sampling

The up-conversion process, i.e., from a 4:2:0 representation back to a 4:4:4 representation, is discussed in clause 10.3.

7.2.4 Floating-point to fixed-point (narrow range) 10 bit conversion

A key component of the pre-encoding process is the conversion from a floating-point representation to a fixed-point, narrow range, 10-bit representation. This process is essentially a quantization step that would introduce some distortion. In general, the conversion process can be expressed as:

$$\mathrm{D}' = \mathrm{Clip3}\left(0, 2^{b} - 1, \mathrm{Round}\left((\mathrm{E}' * \ \mathrm{scale} \ + \ \mathrm{offset})\right)\right) \quad (7\text{-}23)$$

or equivalently:

$$\mathrm{D}' = \mathrm{Clip3}\left(0, (1 \ll b) - 1, \mathrm{Round}\left((\mathrm{E}' * \mathrm{scale} \ + \ \mathrm{offset})\right)\right) \quad (7\text{-}24)$$

where E′ is the floating-point representation of a particular component and D′ is the resulting quantized value using b bits. In this document, b = 10. The scale and offset constants depend on the target range (narrow versus full range video) and the component type (luma, chroma, or colour primary components). More specifically, for the narrow range NCL representation, the scale and offset for the luma component are set as:

$$\mathrm{scale} = 219 * 2^{b-8} = 219 * \left(1 \ll (b - 8)\right) \quad (7\text{-}25)$$

$$\mathrm{offset} = 2^{b-4} = 1 \ll (b - 4) \quad (7\text{-}26)$$

$$\mathrm{DY}' = \mathrm{Clip3}\left(0, \ (1 \ll b) - 1, \mathrm{Round}\left(\left(\mathrm{Y}' * \ \ 219 * \left(1 \ll (b - 8)\right) + \ \ 1 \ll (b - 4)\right)\right)\right) \quad (7\text{-}27)$$

On the other hand, the fixed-point narrow range representation for the two chroma components can be computed as follows:

$$\mathrm{scale} = 224 * 2^{b-8} = 224 * \left(1 \ll (b - 8)\right) \quad (7\text{-}28)$$

$$\mathrm{offset} = 2^{b-1} = 1 \ll (b - 1) \quad (7\text{-}29)$$

and

$$\mathrm{DCb} = \mathrm{Clip3}\left(0,\ (1 \ll b) - 1, \mathrm{Round}\left(\left(\mathrm{Cb} * \ \ 224 * (1 \ll (b-8)) + \ \ 1 \ll (b-1)\right)\right)\right) \tag{7-30}$$

$$\mathrm{DCr} = \mathrm{Clip3}\left(0,\ (1 \ll b) - 1, \mathrm{Round}\left(\left(\mathrm{Cr} * \ \ 224 * (1 \ll (b-8)) + \ \ 1 \ll (b-1)\right)\right)\right) \tag{7-31}$$

NOTE – For a 10 bit narrow range representation, DY′ results in a value within the range of [64, 940]. Similarly, DCb and DCr result in values within the range of [64, 960].

Figure 6 presents the mapping of non-normalized, grey (R=G=B) linear light values to a non-linear representation according to certain transfer functions; HDR (PQ) and SDR (gamma 2.4 for [ITU-R BT.709] / [ITU-R BT.2020], assuming the use of the ITU-R BT.1886 EOTF during display) specifications.

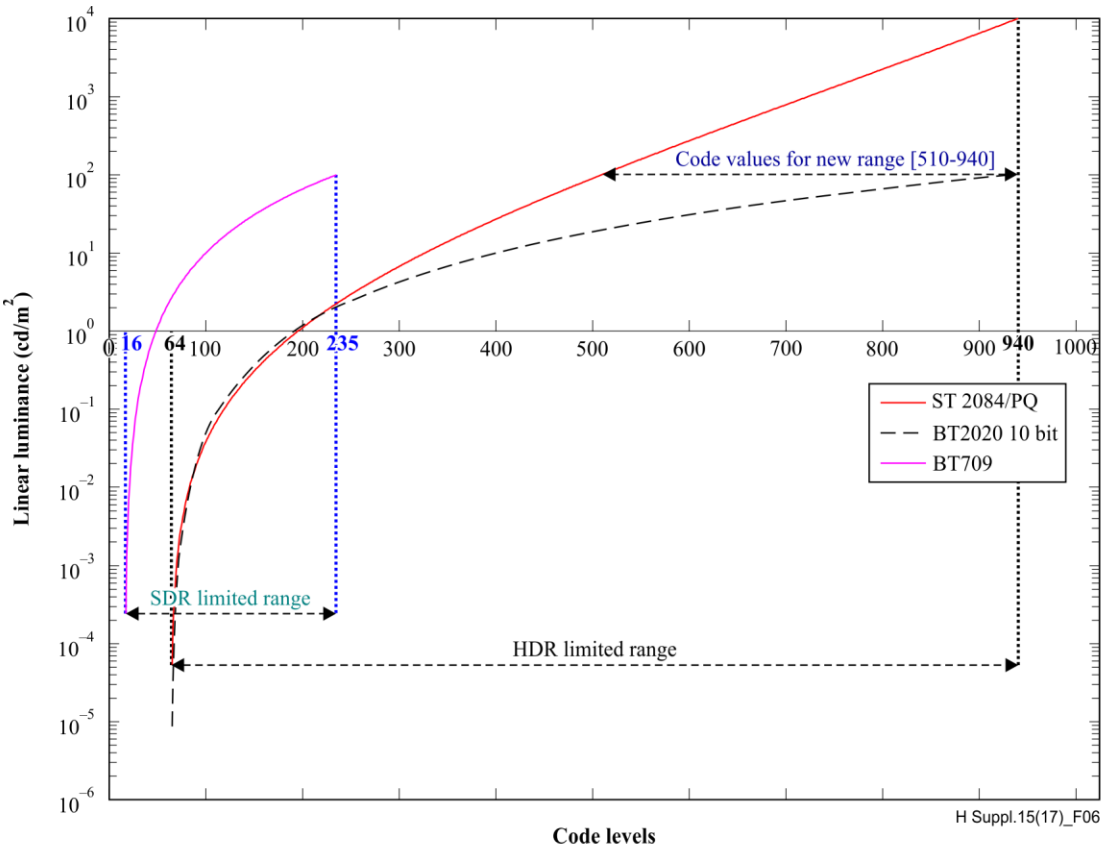

Figure 6 – Mapping of "grey" linear light values to quantized 8 (SDR only) and 10 bit (SDR and HDR) values

The inverse conversion process, i.e., converting from a fixed-point representation back to a floating-point representation, is discussed in detail in clause 10.2.

7.3 Closed loop pre-encoding conversion – luma adjustment

7.3.1 General

As mentioned in clause 7.2.3, chroma leakage may occur in the NCL representation, primarily due to chroma down-sampling, potentially resulting in objectionable artefacts.

This clause presents an alternative conversion method, which can considerably alleviate this problem. This method, is called the luma adjustment method, and it is basically a closed loop conversion process where the impact of chroma down-sampling, quantization, inverse quantization and up-sampling, is accounted for during the luma conversion process. An example schematic diagram of such a system is presented in Figure 7. An iterative luma adjustment method, which is used in the enhanced reference model, is presented in clause 7.3.2. A closed form approach is presented in clause 7.3.3 that requires no iterations and is considerably faster than the iterative method.

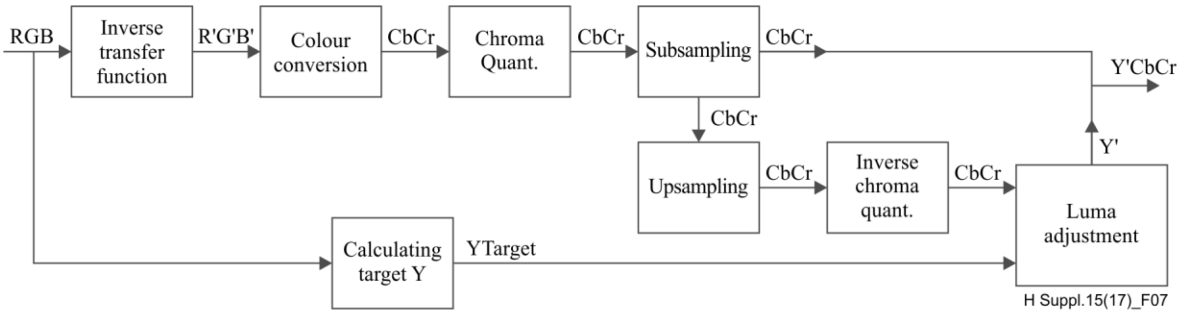

Figure 7 – Example schematic diagram of a closed loop pre-encoding conversion system

7.3.2 Luma adjustment – iterative approach

7.3.2.1 General

Clause 7.2.2 and Formula 7-10 presented a conversion process from R′G′B′ to the Y′CbCr NCL representation. Given this process, the Cb and Cr components can be computed as follows:

$$\mathrm{Cb} = \mathrm{w_{CbR}} * \mathrm{R'} + \mathrm{w_{CbG}} * \mathrm{G'} + \mathrm{w_{CbB}} * \mathrm{B'} \quad (7\text{-}32)$$

$$\mathrm{Cr} = \mathrm{w_{CrR}} * \mathrm{R'} + \mathrm{w_{CrG}} * \mathrm{G'} + \mathrm{w_{CrB}} * \mathrm{B'} \quad (7\text{-}33)$$

Converting these two components to their target resolution, i.e., using the steps described in clause 7.2.3, followed by conversion back to the original representation resolution, as presented in clause 10.3, provides the opportunity to analyse the error introduced into the signal and potentially compensate for it. More specifically, performing quantization, down-scaling (QD), and subsequently up-scaling and inverse quantization (IQU) onto these components would result into the reconstructed $\widetilde{\mathrm{Cb}}$ and $\widetilde{\mathrm{Cr}}$ components, which are defined as follows:

$$\widetilde{\mathrm{Cb}} = \mathrm{IQU}\big(\mathrm{QD}(\mathrm{Cb})\big) \quad (7\text{-}34)$$

$$\widetilde{\mathrm{Cr}} = \mathrm{IQU}\big(\mathrm{QD}(\mathrm{Cr})\big) \quad (7\text{-}35)$$

Luminance (Y), unlike luma (Y′), is computed given the linear R, G and B component values using a formulation of the form:

$$\mathrm{Y} = w_{YR} * \mathrm{R} + w_{YG} * \mathrm{G} + w_{YB} * \mathrm{B} \quad (7\text{-}36)$$

Since $\mathrm{R} = \mathrm{EOTF}(\mathrm{R'})$, this can be rewritten as:

$$\mathrm{Y} = w_{YR} * \mathrm{EOTF}(\mathrm{R'}) + w_{YG} * \mathrm{EOTF}(\mathrm{G'}) + w_{YB} * \mathrm{EOTF}(\mathrm{B'}) \quad (7\text{-}37)$$

However, since the reconstructed $\widetilde{\mathrm{Cb}}$ and $\widetilde{\mathrm{Cr}}$ values will likely differ from the original values Cb and Cr values, the reconstructed $\widetilde{\mathrm{R}}'$, $\widetilde{\mathrm{G}}'$ and $\widetilde{\mathrm{B}}'$ values will also differ from the original R′, G′ and B′ values. Therefore, the reconstructed luminance $\mathrm{Y_{rec}}$, which is equal to:

$$\mathrm{Y_{rec}} = w_{YR} * \mathrm{EOTF}\big(\widetilde{\mathrm{R}}'\big) + w_{YG} * \mathrm{EOTF}\big(\widetilde{\mathrm{G}}'\big) + w_{YB} * \mathrm{EOTF}\big(\widetilde{\mathrm{B}}'\big) \quad (7\text{-}38)$$

will also differ from the original luminance Y. Using Formula 10-2 it can be computed that:

$$\mathrm{Y}_{rec} = w_{YR} * \mathrm{EOTF}\big(\mathrm{Y'} + a_{RCr} * \widetilde{\mathrm{Cr}}\big) + w_{YG} * \mathrm{EOTF}\big(\mathrm{Y'} + a_{GCb} * \widetilde{\mathrm{Cb}} + a_{GCr} * \widetilde{\mathrm{Cr}}\big)$$
$$+ \ w_{YB} * \mathrm{EOTF}\big(\mathrm{Y'} + a_{BCb} * \widetilde{\mathrm{Cb}}\big) \quad (7\text{-}39)$$

Chroma component dependent factors can then be defined as follows:

$$\mathrm{Crfactor} = a_{RCr} * \widetilde{\mathrm{Cr}} \quad (7\text{-}40)$$

$$\mathrm{Gfactor} = a_{GCb} * \widetilde{\mathrm{Cb}} + a_{GCr} * \widetilde{\mathrm{Cr}} \quad (7\text{-}41)$$

$$\mathrm{Cbfactor} = a_{BCb} * \widetilde{\mathrm{Cb}} \quad (7\text{-}42)$$

resulting in the following formulation for Y_{rec}:

$$Y_{\mathrm{rec}} = w_{YR} * \mathrm{EOTF}(Y' + \mathrm{Crfactor}) + w_{YG} * \mathrm{EOTF}(Y' + \mathrm{Gfactor}) + w_{YB} * \mathrm{EOTF}(Y' + \mathrm{Cbfactor}) \quad (7\text{-}43)$$

The intent of the luma adjustment method is to try and locate the value of Y' that would minimize distortion for Y_{rec} compared to the original luminance value Y. Unfortunately, due to the non-linear characteristics of the EOTF, solving for Y' given Y and the values of Crfactor, Gfactor and Cbfactor is not a straightforward process. However, root-finding numerical methods, such as the bisection method, can be used instead. The performance, and more specifically the convergence speed and accuracy of these methods, is considerably impacted by the selection of the initial interval as well as the computations performed during the search.

NOTE – Alternative methods that try to approximate the impact on Y or methods that evaluate the impact on other components have also been suggested and have remained under study.

The target of the luma adjustment process is to minimize luminance distortion, which can be realized through the following ordered steps:

a) Calculate the luminance value Y from the original R, G and B, e.g., using Formula 7-36. This will be referred to as the Y_{target} value.

b) Convert the R, G and B data to their R', G' and B' representation.

c) Given the R', G' and B' planes generate the Cb and Cr chroma planes.

d) Down-scale and quantize the chroma planes to their target representation.

e) De-quantize and up-convert the chroma planes back to their original representation, i.e., $\widetilde{\mathrm{Cb}}$ and $\widetilde{\mathrm{Cr}}$.

f) Calculate Crfactor, Gfactor and Cbfactor from $\widetilde{\mathrm{Cb}}$ and $\widetilde{\mathrm{Cr}}$.

g) Given the reconstructed chroma planes, try to find for each luma position an appropriate Y' value, i.e., Y'_{adjust}, that would potentially result in a minimum distortion for a particular aspect of the signal, i.e., in this case minimum luminance distortion. The value of Y'_{adjust} at each luma position would be the value used for the encoding of the luma signal. In particular, Y'_{adjust} would be the solution to the equation:

$$\begin{aligned} Y_{\mathrm{target}} = {} & w_{YR} * \mathrm{EOTF}\left(Y'_{\mathrm{adjust}} + \mathrm{Crfactor}\right) + w_{YG} * \mathrm{EOTF}\left(Y'_{\mathrm{adjust}} + \mathrm{Gfactor}\right) \\ & + w_{YB} * \mathrm{EOTF}\left(Y'_{\mathrm{adjust}} + \mathrm{Cbfactor}\right) \end{aligned} \quad (7\text{-}44)$$

7.3.2.2 Bisection search

The bisection search method is an iterative technique that is commonly used to derive the roots of an equation of the form $\mathrm{f}(x) = 0$. The function $\mathrm{f}(x)$ is assumed to be continuous and defined over an interval [a, b], where $\mathrm{f(a)}$ and $\mathrm{f(b)}$ need to have opposite signs. In this application, it is desirable to have a unique solution. For this to be guaranteed, the behaviour of $\mathrm{f}(x)$ within this interval needs to also be strictly monotonic (i.e., consistently increasing or decreasing). At each iteration of the search the interval is divided by two, i.e., it is divided at the midpoint $c = \frac{(a+b)}{2}$. Then the value of the function $\mathrm{f(c)}$ is computed at this point. Depending on the value of $\mathrm{f(c)}$ and its relationship with $\mathrm{f(a)}$ and $\mathrm{f(b)}$, a new smaller interval is defined that satisfies the opposite sign condition. This search is repeated until a root is found, the interval is sufficiently small, or if a certain maximum number of iterations has been achieved.

This method can be used to find the Y'_{adjust} value for luma, as discussed in the previous clause, in the following way:

Let x be the value that represents the quantized representation of the luma component, as defined in clause 7.2.1. Furthermore, let $\mathrm{g}(x)$ be the dequantization function (clause 10.5) that maps the value of x back to its original representation and essentially a value between [0, 1]. Then, the narrow range, 10-bit representation can be computed as:

$$\mathrm{g}(x) = (x - 64.0) \div 876.0. \quad (7\text{-}45)$$

Now let $\mathrm{f}(x)$ be the function:

$$\begin{aligned} \mathrm{f}(x) = {} & w_{YR} * \mathrm{EOTF}(\mathrm{g}(x) + \mathrm{Crfactor}) + w_{YG} * \mathrm{EOTF}(\mathrm{g}(x) + \mathrm{Gfactor}) \\ & + w_{YB} * \mathrm{EOTF}(\mathrm{g}(x) + \mathrm{Cbfactor}) - Y_{target} \end{aligned} \quad (7\text{-}46)$$

The initial interval can be set as the entire range, i.e., [a, b] = [64, 940]. Given the characteristics of the PQ EOTF it is expected that $\mathrm{f(a)} \leq 0$ and $\mathrm{f(b)} \geq 0$, which satisfy the bisection conditions for a unique solution. The value of $\mathrm{f}(x)$ at position x can be computed as $x = \frac{(a+b)}{2} = 502$ and the interval can then be adjusted accordingly. If, for example, $\mathrm{f}(502) > 0$ then the interval will be adjusted to [64,502]. The next evaluation point will then be the middle point of this new interval, i.e., the point at $x = \frac{(64+502)}{2} = 283$. At this point, if now $\mathrm{f}(283) < 0$ then the interval will be adjusted to [283, 502]. The process can continue until either a value is found that satisfies $\mathrm{f}(x) = 0$, or when the interval is of the form [k, k+1], e.g., [343, 344]. In this case both values can be evaluated and the one resulting in the smallest distortion for Y or $\mathrm{EOTF}^{-1}(\mathrm{Y})$ can be used.

A critical component of this method is the selection of the initial interval. The brute force approach is to use the entire valid range as the initial range, e.g., for the 10-bit narrow representation the range of [64, 940] as in the previous example. This, though, may require, in the worst case, a number of iterations equal to the target bit depth of the content to reach an interval of size one. However, using information about the original colour and the chroma values Cb^* and Cr^*, upper and lower bounds can be found for the value of $\mathrm{Y}'_{\mathrm{adjust}}$, greatly reducing the size of the initial interval. This can considerably reduce the average number of iterations and thus have a direct impact on the number of computations performed and the overall complexity of the process as described in [Strom PCS2016]. Three such bounds are preferably used. The first is described in [Strom DCC2016] and uses the following definitions:

$$\mathrm{R}'_{\mathrm{bound}} = \mathrm{EOTF}^{-1}\left(\mathrm{Y}_{\mathrm{target}}\right) - \mathrm{Crfactor} \quad (7\text{-}47)$$

$$\mathrm{G}'_{\mathrm{bound}} = \mathrm{EOTF}^{-1}\left(\mathrm{Y}_{\mathrm{target}}\right) - \mathrm{Gfactor} \quad (7\text{-}48)$$

$$\mathrm{B}'_{\mathrm{bound}} = \mathrm{EOTF}^{-1}\left(\mathrm{Y}_{\mathrm{target}}\right) - \mathrm{Cbfactor}. \quad (7\text{-}49)$$

Y' will always be in the interval $[\mathrm{Min}(\mathrm{R}'_{\mathrm{bound}}, \mathrm{G}'_{\mathrm{bound}}, \mathrm{B}'_{\mathrm{bound}}), \mathrm{Max}(\mathrm{R}'_{\mathrm{bound}}, \mathrm{G}'_{\mathrm{bound}}, \mathrm{B}'_{\mathrm{bound}})]$. The proof for this is out of the scope of this document, but is presented in [Strom DCC2016]. The second bound uses the fact that all three variables $\mathrm{R}'_{\mathrm{bound}}$, $\mathrm{G}'_{\mathrm{bound}}$ and $\mathrm{B}'_{\mathrm{bound}}$ are smaller than 1, which leads to a tighter upper bound, i.e., $\mathrm{Y}' \leq \mathrm{EOTF}^{-1}(\mathrm{Y}_{target})$. This makes use of the fact that the EOTF is convex as described in [Strom DCC2016]. The third bound relies on the fact that all three of the reconstructed colour components $\widetilde{\mathrm{R}}'$, $\widetilde{\mathrm{G}}'$ and $\widetilde{\mathrm{B}}'$ cannot simultaneously be smaller (or all simultaneously larger) than the original colour components R', G' and B' for the computation of $\mathrm{Y}'_{\mathrm{adjust}}$. By using the following definitions:

$$\mathrm{Y}'_{\widetilde{\mathrm{Cb}}} = \mathrm{R}' + \mathrm{Crfactor} \quad (7\text{-}50)$$

$$\mathrm{Y}'_{\mathrm{G}} = \mathrm{G}' + \mathrm{Gfactor} \quad (7\text{-}51)$$

$$\mathrm{Y}'_{\widetilde{\mathrm{Cr}}} = \mathrm{B}' + \mathrm{Cbfactor} \quad (7\text{-}52)$$

it can be shown that $\mathrm{Y}'_{\mathrm{adjust}}$ must be in the interval:

$$[\mathrm{Y}'_{\mathrm{min}}, \mathrm{Y}'_{\mathrm{max}}] = \left[\mathrm{Min}\left(\mathrm{Y}'_{\widetilde{\mathrm{Cb}}}, \mathrm{Y}'_{\mathrm{G}}, \mathrm{Y}'_{\widetilde{\mathrm{Cr}}}\right), \mathrm{Max}\left(\mathrm{Y}'_{\widetilde{\mathrm{Cb}}}, \mathrm{Y}'_{\mathrm{G}}, \mathrm{Y}'_{\widetilde{\mathrm{Cr}}}\right)\right] \quad (7\text{-}53)$$

By combining all three bounds together the minimum and maximum bounds

for Y' can be computed as:

$$\mathrm{Y}'_{\mathrm{low\_bound}} = \mathrm{Max}(\mathrm{Min}(\mathrm{R}'_{\mathrm{bound}}, \mathrm{G}'_{\mathrm{bound}}, \mathrm{B}'_{\mathrm{bound}}), \mathrm{Y}'_{\mathrm{min}}) \quad (7\text{-}54)$$

$$\mathrm{Y}'_{\mathrm{high\_bound}} = \begin{cases} \mathrm{Min}\left(\mathrm{EOTF}^{-1}\left(\mathrm{Y}_{\mathrm{target}}\right), \mathrm{Y}'_{\mathrm{max}}\right) & ; \text{ if } \mathrm{Max}(\mathrm{R}'_{\mathrm{bound}}, \mathrm{G}'_{\mathrm{bound}}, \mathrm{B}'_{\mathrm{bound}}) < 1 \\ \mathrm{Min}(\mathrm{Max}(\mathrm{R}'_{\mathrm{bound}}, \mathrm{G}'_{\mathrm{bound}}, \mathrm{B}'_{\mathrm{bound}}), \mathrm{Y}'_{\mathrm{max}}) & ; \text{ otherwise} \end{cases} \quad (7\text{-}55)$$

Finally, the initial interval [a, b] for x is calculated as $\mathrm{a} = \mathrm{Floor}\left(\mathrm{g}^{-1}\left(\mathrm{Y}'_{\mathrm{low\_bound}}\right)\right)$ and $\mathrm{b} = \mathrm{Ceil}\left(\mathrm{g}^{-1}\left(\mathrm{Y}'_{\mathrm{high\_bound}}\right)\right)$ where $\mathrm{g}^{-1}(x)$ is the inverse of Formula 7-45.

(a) Originals in 4:4:4 (b) Traditional subsampling (c) Luma adjustment based subsampling

Figure 8 – Tone-mapped examples showing improvements of the luma adjustment method

Figure 8 shows an example of what the difference can be when performing traditional subsampling, according to the simple reference model as described in clause 7.2.3 and using the luma component unchanged, (Figure 8 (b)) compared to the luma adjustment method that is described in this clause (Figure 8 (c)) and is used in the enhanced reference model. These images were processed using the Rec. ITU-R BT.709 colour primaries to more easily demonstrate the differences. Since the printed medium cannot reproduce HDR images, tone-mapped versions are calculated using

$$\mathrm{R_{SDR}} = \mathrm{Clip3}\left(0, 255, 255 * (\mathrm{R_{HDR}} * 2^{c})^{\frac{1}{\gamma}}\right), \tag{7-56}$$

where γ = 2.2 and c is an exposure parameter set to make the SDR image look similar to the HDR image. The artefacts are clearly more visible in the simple reference model subsampling case compared to when using the enhanced reference model method.

7.3.3 Luma adjustment – closed form solution

The bisection search method described in the previous clause, has a worst-case complexity for a 10 bit data representation of ten iterations. This might be a problem for certain real-time applications and, in particular, for hardware implementations. Quite often, hardware systems are designed by taking into account the worst-case scenario and by assuming that the same number of processing steps is required for each block or sample. Even though the complexity of the bisection method could be further bounded by limiting the maximum number of iterations, it may be beneficial for such applications to use a closed-form solution that is able to determine an appropriate luma value in a single step.

A closed-form solution can be found as follows. First, down-sampled chroma samples are obtained. Then, chroma is up-sampled to the original (luma) resolution by applying a chosen up-sampling filter. Then, for every pixel, the algorithm estimates a luma value $\mathrm{Y'_{adjust}}$ that, in combination with the up-sampled $\widetilde{\mathrm{Cb}}$ and $\widetilde{\mathrm{Cr}}$ values, will result in a reconstructed $\widetilde{\mathrm{RGB}}$ pixel with colour component values of $\{\widetilde{\mathrm{R}}, \widetilde{\mathrm{G}}, \widetilde{\mathrm{B}}\}$. It is highly desirable that $\widetilde{\mathrm{RGB}}$ is as close as possible to the original linear light RGB value according to a chosen distance metric. The difference between the two RGB values is denoted as:

$$D = \mathrm{Diff}\left(\mathrm{RGB}\ \ , \widetilde{\mathrm{RGB}}\right) \tag{7-57}$$

Depending on the chosen distance metric $\mathrm{Diff}(x)$, different closed-form solutions to the optimization problem can be obtained. In the following, a solution based on the weighted sum of the differences of linear R, G and B data is described.

In particular, the square of the sum of the weighted differences between the individual R, G and B components can be computed as:

$$D = \left(\mathrm{w_R} * \left(\tilde{\mathrm{R}} - \mathrm{R}\right) + \mathrm{w_G} * \left(\tilde{\mathrm{G}} - \mathrm{G}\right) + \mathrm{w_B} * \left(\tilde{\mathrm{B}} - \mathrm{B}\right)\right)^2. \tag{7-58}$$

This is also equivalent to the formula

$$D = \left(\mathrm{w_R} * \left(\mathrm{EOTF}(\tilde{\mathrm{R}}') - \mathrm{EOTF}(\mathrm{R}')\right) + \mathrm{w_G} * \left(\mathrm{EOTF}(\tilde{\mathrm{G}}') - \mathrm{EOTF}(\mathrm{G}')\right) + \mathrm{w_B} * \left(\mathrm{EOTF}(\tilde{\mathrm{B}}') - \mathrm{EOTF}(\mathrm{B}')\right)\right)^2, \tag{7-59}$$

where $\mathrm{EOTF}(x)$ is the PQ EOTF. If weights $\mathrm{w_R}$, $\mathrm{w_G}$ and $\mathrm{w_B}$ are set equal to the contribution of the linear R, G and B components to the luminance component, this cost function would be minimizing the squared difference between the luminance values.

NOTE 1 – It is also possible to use other error functions, such as the sum of the R, G and B component squared errors, which may result in different solutions.

Finding a closed-form solution for Y′ may be difficult because of the non-trivial form of the PQ EOTF. In order to obtain a closed form solution, the $\mathrm{EOTF(x)}$ is approximated with a first-degree polynomial using the truncated Taylor series expansion:

$$\mathrm{EOTF}(x_i + \Delta) \approx \mathrm{EOTF}(x_i) + \mathrm{EOTF}'(x_i) * \Delta_x, \tag{7-60}$$

where $\mathrm{EOTF}'(x_i)$ is the value of the derivative of $\mathrm{EOTF}(x)$ with respect to x at point x_i and Δ_X is the change in the value of x. Substituting Formula 7-60 into Formula 7-59, the cost function is approximated as follows:

$$D = (\mathrm{w_R} * \mathrm{EOTF}'(\mathrm{R}') * \Delta_R + \mathrm{w_G} * \mathrm{EOTF}'(\mathrm{G}') * \Delta_G + \mathrm{w_B} * \mathrm{EOTF}'(\mathrm{B}') * \Delta_B)^2. \tag{7-61}$$

Colour component values $\tilde{\mathrm{R}}'$, $\tilde{\mathrm{G}}$ and $\tilde{\mathrm{B}}'$ in the EOTF domain can be obtained from Y′, $\widetilde{\mathrm{Cb}}$ and $\widetilde{\mathrm{Cr}}$ data using the inverse colour transformation described in clause 10.4 using Formula 10-2.

Substituting Formula 10-2 into Formula 7-61, results in

$$D = (\mathrm{w_R} * \mathrm{EOTF}'(\mathrm{R}') * \left(a_{RY} * \mathrm{Y}'_{\mathrm{adjust}} + a_{RCb} * \widetilde{\mathrm{Cb}} + a_{RCr} * \widetilde{\mathrm{Cr}} - (a_{RY} * \mathrm{Y}' + a_{RCb} * \mathrm{Cb} + a_{RCr} * \mathrm{Cr})\right) +$$

$$\mathrm{w_G} * \mathrm{EOTF}'(\mathrm{G}') * \left(a_{GY} * \mathrm{Y}'_{\mathrm{adjust}} + a_{GCb} * \widetilde{\mathrm{Cb}} + a_{GCr} * \widetilde{\mathrm{Cr}} - (a_{GY} * \mathrm{Y}' + a_{GCb} * \mathrm{Cb} + a_{GCr} * \mathrm{Cr})\right) +$$

$$\mathrm{w_B} * \mathrm{EOTF}'(\mathrm{B}') * \left(a_{BY} * \mathrm{Y}'_{\mathrm{adjust}} + a_{BCb} * \widetilde{\mathrm{Cb}} + a_{BCr} * \widetilde{\mathrm{Cr}} - (a_{BY} * \mathrm{Y}' + a_{BCb} * \mathrm{Cb} + a_{BCr} * \mathrm{Cr})\right))^2 \tag{7-62}$$

Sorting the expressions inside the brackets and substituting with their numerical values those coefficients in matrix **A** from Formula 10-2 that are equal to 0 and 1 results in

$$D = \left(\mathrm{w_R} * \mathrm{EOTF}'(\mathrm{R}') * \left(\mathrm{Y}'_{\mathrm{adjust}} - e_R\right) + \mathrm{w_G} * \mathrm{EOTF}'(\mathrm{G}') * \left(\mathrm{Y}'_{\mathrm{adjust}} - e_G\right) + \mathrm{w_B} * \mathrm{EOTF}'(\mathrm{B}') * \left(\mathrm{Y}'_{\mathrm{adjust}} - e_B\right)\right)^2, \tag{7-63}$$

where e_R, e_G and e_B are defined as follows

$$e_R = \mathrm{Y}' - \left(\widetilde{\mathrm{Cr}} - \mathrm{Cr}\right) * a_{\mathrm{RCr}}, \tag{7-64}$$

$$e_G = \mathrm{Y}' - \left(\widetilde{\mathrm{Cb}} - \mathrm{Cb}\right) * a_{\mathrm{GCb}} - \left(\widetilde{\mathrm{Cr}} - \mathrm{Cr}\right) * a_{\mathrm{GCr}}, \tag{7-65}$$

$$e_B = \mathrm{Y}' - \left(\widetilde{\mathrm{Cb}} - \mathrm{Cb}\right) * a_{\mathrm{BCb}}. \tag{7-66}$$

Then, in order to find the local minimum, D is differentiated with respect to $\mathrm{Y}'_{\mathrm{adjust}}$ (e_R, e_G and e_B do not depend on $\mathrm{Y}'_{\mathrm{adjust}}$), the derivative is set equal to zero, and the resulting equation is solved with respect to Y′. The value of Y′ is then equal to:

$$Y'_{adjust} = \frac{w_R * EOTF'(R') * e_R + w_G * EOTF'(G') * e_G + w_B * EOTF'(B') * e_B}{w_R * EOTF'(R') + w_G * EOTF'(G') + w_B * EOTF'(B')} \quad (7\text{-}67)$$

Applying Formulae 7-64, 7-65, 7-66 and 7-67, one can obtain the adjusted value of Y′ in a single step. This approach can be adopted by applications that would benefit from or require lower complexity and a fixed number of operations per luma sample. Experimental results suggest that the algorithm described above can achieve a considerable complexity reduction, while at the same time having only a small difference in terms of objective metric performance compared to the bisection search method. In a particular reference implementation, a speed-up factor of around 2.5 times for the total colour conversion runtime compared to the bisection search method was reported in [Norkin2016]. The differences in the objective measurements are mostly due to the approximation of the EOTF with its tangent. Subjective performance appears to be similar to that of the bisection search method.

The derivative $EOTF'(x)$ in Formula 7-67 can be computed using a formula obtained by the differentiation of the EOTF or by the definition of a derivative, i.e., by dividing the change in the function value by the increment of the function argument. Alternatively, the derivative values can be pre-computed and stored in an LUT.

As mentioned earlier, the weights w_R, w_G and w_B can be chosen based on the desired precision or importance of each component. For example, they can be set equal to 1 or based on the contribution of each colour component to the luminance CIE 1931 Y component.

The above algorithm can be summarized as follows:

a) Convert the original R, G and B data to their R′, G′ and B′ representation, if needed.

b) Given the R′, G′ and B′ planes generate the Cb and Cr chroma planes.

c) Down-scale chroma planes to 4:2:0 or 4:2:2 representation and quantize the samples.

d) De-quantize and up-convert the chroma back to their original resolution to obtain $\widetilde{Cb}$ and $\widetilde{Cr}$.

e) For each luma sample, calculate e_R, e_G and e_B based on Formulae 7-64, 7-65 and 7-66, respectively.

f) Calculate Y'_{adjust} based on Formula 7-67.

NOTE 2 – The formulae described above do not take into account the effects of clipping of the R, G and B data, within the range of 0 to 10 000 cd/m^2, when applying the colour transformation from Y′CbCr to RGB. This may decrease the precision of the Y'_{adjust} estimation obtained from Formula 7-67 when R, G and B values are close to their upper limit of 10 000 cd/m^2. This clipping effect can be mitigated by modifying Formula 7-67 when one or more of the R, G and B samples are clipped at 10 000 cd/m^2. The details of such a modification are considered as being out of the scope of this document. It can be argued that this effect would not be significant for most of the currently available HDR/WCG content given that, due to limitations of existing displays, HDR/WCG content are rarely mastered with a peak luminance value close to 10 000 cd/m^2. Therefore, the results obtained with this solution are likely difficult to distinguish from the results generated using the bisection search.

NOTE 3 – Several other methods for performing the luma adjustment process using a closed form process, such as methods involving look-up tables, have also been suggested.

8 Encoding process

8.1 General

After preprocessing, the data is ready for compression. The HDR/WCG data coming out of the preprocessing step will exhibit slightly different characteristics than typical, standard dynamic range (SDR) data. This means that it may be possible to increase perceptual/subjective quality if the encoder is configured in a slightly different manner compared to when compressing SDR data. This clause presents two such differences in data characteristics and gives guidance on how an encoder may be configured to better exploit these differences.

8.2 Perceptual luma quantization

8.2.1 General

When processing SDR data, a power law transfer function such as the one described in [ITU-R BT.709] is typically used. As is described above, the HDR/WCG data has instead undergone processing using the PQ transfer function defined in [SMPTE ST 2084] and [ITU-R BT.2100]. This will in itself give a different characteristic of the processed data. One way to see this is to preprocess the same SDR data using both the ITU-R BT.709 transfer function and the PQ transfer function. For a 10-bit representation, if the original data has a peak brightness of 100 cd/m^2, the luma component will occupy all code levels from 64 to 940 if the ITU-R BT.709 transfer function is used. However, only code levels from 64 to 509 will be used in the case of the PQ transfer function. Since the step sizes are different in the two cases, a perturbation of +/− 1 code level around code level 509 (100 cd/m^2) in the PQ case will be the equivalent of roughly +/− 4 code levels around code level 940 (also 100 cd/m^2) in the ITU-R BT.709 case. At the same time, a perturbation of +/− 1 code level around code level 80 (0.01 cd/m^2) in the PQ case will be roughly equivalent to a perturbation of +/− 1 code level around code level 80 (0.01 cd/m^2) in the ITU-R BT.709 case. Thus, if an encoder is wired to treat an error of one code level the same way regardless if it is at level 80 or 509, it will allow errors that are four times larger in the bright areas (at around 100 cd/m^2) if it uses the PQ transfer function compared to the use of [ITU-R BT.709]. In other words, by switching from an ITU-R BT.709 transfer function to PQ, a lot of bits will be redistributed from the bright areas of the image to the dark areas.

Thus, if an encoder with a certain setting has achieved a good balance between bright and dark areas for the ITU-R BT.709 transfer function, using the encoder with the same settings for PQ may produce images in which bright and dark areas are allocated too few and too many bits, respectively. This may result in more objectionable compression artefacts in the bright areas, while no perceivable improvement may be observed in the dark areas. For HDR/WCG data this effect can be even more pronounced; the luminance increase from a code level increase is even higher at, for example, 4 000 cd/m^2 than it is at 100 cd/m^2. Furthermore, it might be the case that the HDR/WCG content contains considerable amounts of noise in the dark areas, which may have a further impact on performance.

One way to ameliorate this effect is for the encoder to calculate the average luma value in a block and, using this value, adaptively adjust the block's quantization parameter (QP). In particular, an encoder may increase or decrease the QP for the block if it is classified as a *dark* or a *bright* block respectively. In this way, it may be possible to shift bits back from dark regions to bright regions and potentially achieve a result that may be perceptually more pleasing.

Shifting bits from dark to bright areas works in the opposite direction of the inverse EOTF, which assigns more code levels to dark values. However, the inverse EOTF is based on the best-case sensitivity for the human visual system. For instance, if all colours in a picture are dark, it predicts well how a small perturbation can be detected. However, if some colours are dark and others bright, it is harder for the visual system to detect perturbations in the dark areas, and hence it is reasonable to move bits from dark to bright areas.

NOTE – In the development of this technique, only the local luma characteristics were analysed, without trying to adapt performance based on regional or global brightness characteristics, among other potential considerations. Other aspects, such as the noise present in some of the material, may have also impacted coding performance and affected the design of the scheme. Further study may result in a revision of the described methods if additional evidence on the behaviour of the technology is obtained.

Many existing coders already use some form of adaptive QP method. As an example, such methods can be used to increase the QP in areas of very high variance (where it is perceptually hard to see errors) and decrease the QP in areas of lower variance (where errors are typically more visible). In some other systems, brightness, edges, motion, as well as other features, may also be considered. However, these methods are likely to have been designed based on SDR content characteristics. Given the above observations regarding the transfer function relationships, it is advised that, when compressing PQ encoded data, a QP adaptation method is considered that also takes into account these relationships. Other characteristics, such as colour, could also be considered.

A simple example QP adaptation method, which is used in the enhanced reference model, is presented below. This method was found to result in better subjective as well as objective performance compared to the fixed QP coding configuration that is used in the simple reference model.

8.2.2 Luma-dependent adaptive quantization – an example

The purpose of this approach is to try and match a similar level of distortion to a particular, grey level, luminance value x when either the power law transfer function of ITU-R BT.709 $\mathrm{f_{709}(x)}$ or the PQ transfer function $\mathrm{f_{PQ}(x)}$ are used, in combination with 10 bit quantization as well as a codec's quantization level. More specifically, it is highly desirable to determine the QP value QP_{PQ} to be used with a PQ encoded value x, that would result in the same or similar distortion, or equivalently the same or similar quantization behaviour Quant( ), if that same value was encoded using the ITU-R BT.709 transfer function and a known QP value QP_{709}. That is:

$$\mathrm{Quant}(\mathrm{f_{709}(x), QP_{709}}) \cong \mathrm{Quant}\left(\mathrm{f_{PQ}(x), QP_{PQ}}\right) \quad (8\text{-}1)$$

The linear characteristics of the transformations employed on residual data in codecs such as AVC and HEVC enable the consideration of these formulations even after such transformations are performed. However, these also limit the consideration of such an optimization at a block level. Based on the characteristics of the ITU-R BT.709 and PQ transfer function and Formula 8-1, an approximate relationship between QP_{PQ} and QP_{709} can be computed as:

$$\mathrm{QP_{PQ} = QP_{709} + dQP(x)} \quad (8\text{-}2)$$

This relationship is depicted in Table 3, as well as in Figure 9, with int_L replacing the value of x. More specifically, in a particular implementation, int_L is computed by obtaining the average luma value of a 64×64 CTU block, $L_{average}$ and then rounding this quantity, i.e., int_L= Round($L_{average}$). Based on this relationship, for every CTU, the QP will be adjusted according to its brightness by this dQP value.

Table 3 – Look-up table of the dQP value from the average of the luma value

luma int_L range	dQP
$int_L < 301$	3
$301 \le int_L < 367$	2
$367 \le int_L < 434$	1
$434 \le int_L < 501$	0
$501 \le int_L < 567$	−1
$567 \le int_L < 634$	−2
$634 \le int_L < 701$	−3
$701 \le int_L < 767$	−4
$767 \le int_L < 834$	−5
$int_L \ge 834$	−6

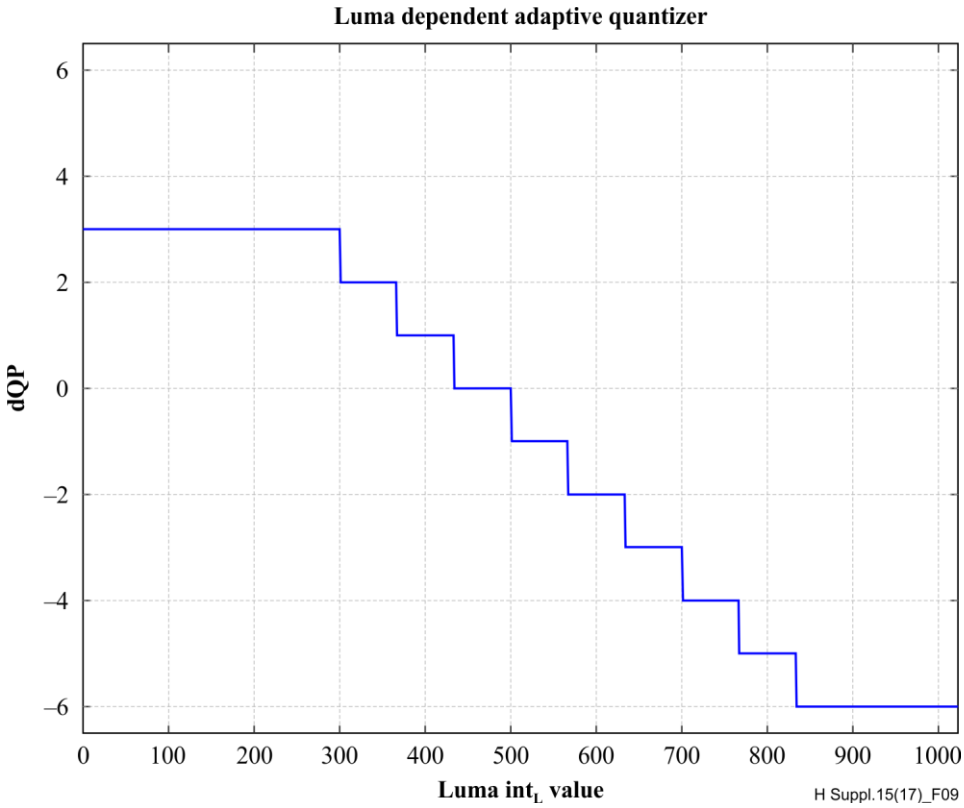

Figure 9 – Difference in QP value as a function of the average luma value in a 64×64 block

8.3 Chroma QP offset

8.3.1 General

Another major difference between HDR/WCG and SDR data has been observed in the characteristics of the chroma channels Cb and Cr. For 10 bit SDR content encoded using the ITU-R BT.709 transfer function and the Rec. ITU-R BT.709 colour space, typically all three components Y, Cb and Cr use the entire allowed range, i.e., Y′ will use up most of the range [64, 940] and Cb and Cr will populate most of [64, 960]. However, for HDR/WCG data using the ITU-R BT.2020 colour space and the PQ transfer function, the Cb and Cr distributions will be clustered closer to the mid-point of 512, which represents a value of Cb and Cr equal to 0. On the other hand, the Y′ component may still populate most of its allowed range. Furthermore, if the content does not exercise the entire ITU-R BT.2020 colour space, the Cb and Cr distributions will be even more tightly clustered around 0. In particular, if SDR content is instead represented using the PQ transfer function and the ITU-R BT.2020 colour space, the distribution of the Cb and Cr will be reduced substantially compared to its original ITU-R BT.709 representation. However, the luminance distribution may not be as affected.

The above observations may have a considerable impact on the encoding process. An existing encoder setting may have been able to achieve a good balance between luma and chroma for SDR content using the ITU-R BT.709 representation. However, the same encoder with the same settings will likely not achieve the same performance for the same content if the content is represented using the PQ transfer function and the ITU-R BT.2020 colour space. Given the characteristics of the new representation, this will result in a bitrate allocation shift from chroma to luma. However, if chroma is not allocated enough bits, this may give rise to visible chroma artefacts. These artefacts may, for example, appear in white areas, where miscolorations in the direction of cyan and magenta can become visible, as seen in Figure 10 (a).

One way to ameliorate this is for the encoder to apply a negative chroma QP offset value. This will lower the QP value used for quantizing the chroma coefficients and has an effect similar to stretching out the Cb and Cr distributions. This effectively shifts bits back from luma to chroma and thus allowing the encoder to achieve a better balance between chroma and luma quality.

Since chroma artefacts typically become more visible at low bit rates, applying a large negative chroma QP offset at such rates can potentially help reduce these artefacts significantly. However, after a certain rate point, chroma quality may be considered as being good enough. At this point it may no longer be necessary to shift bits from luma to chroma. Thus, at higher rates the chroma QP offset can be set to a smaller value or even be set to zero.

A special case occurs when it is known that the content is in a restricted subset of the colour gamut defined by the ITU-R BT.2020 and ITU-R BT.2100 colour primaries. As an example, if a mastering display limited to the P3D65 colour primaries, as defined in [SMPTE RP 431-2], was used to grade the content, then it is likely that the content does not also venture outside of this colour gamut. In this case, it might be known in advance that the chroma values will never go outside a certain interval that is much smaller than the allowed [64, 960] range. Under such circumstances, it may be advantageous to use a larger negative chroma QP offset compared to the QP offset that may be used for content that makes use of the entire colour gamut defined by the ITU-R BT.2020 and ITU-R BT.2100 colour primaries.

8.3.2 Example of chroma QP offset settings

In the following example it is assumed that the colour primaries of the mastering display/capture device are known.

Based on this knowledge, a model is used to assign QP offsets for Cb and Cr based on the luma QP and a factor based on the capture and representation colour primaries. The model is expressed as:

$$\mathrm{QPoffsetCb} = \mathrm{Clip3}\left(-12, 0, \mathrm{Round}\left(\mathrm{c_{cb}} * (\mathrm{k} * \mathrm{QP} + \mathrm{l})\right)\right) \tag{8-3}$$

$$\mathrm{QPoffsetCr} = \mathrm{Clip3}\left(-12, 0, \mathrm{Round}\left(\mathrm{c_{cr}} * (\mathrm{k} * \mathrm{QP} + \mathrm{l})\right)\right) \tag{8-4}$$

where c_{cb} = 1 if the capture colour primaries are the same as the representation colour primaries, c_{cb}=1.04 if the capture colour primaries are equal to the P3D65 primaries and the representation colour primaries are equal to the ITU-R BT.2020 primaries, and c_{cb}=1.14 if the capture colour primaries are equal to the ITU-R BT.709 primaries and the representation primaries are equal to the ITU-R BT.2020 primaries.

Similarly, c_{cr} = 1 if the capture colour primaries are the same as the representation colour primaries, c_{cr}=1.39 if the capture colour primaries are equal to the P3D65 primaries and the representation colour primaries are equal to the ITU-R BT.2020 primaries, and c_{cr}=1.78 if the capture colour primaries are equal to the ITU-R BT.709 primaries and the representation primaries are equal to the ITU-R BT.2020 primaries.

Finally, $\mathrm{k} = -0.46$ and $\mathrm{l} = 9.26$.

The constants c_{cr} and c_{cb} have been calculated as the ratio of the range in the different colour representations. As an example, a maximally red colour represented using ITU-R BT.709 primaries is the colour RGB_{709} = (1,0,0). This gives a fully saturated Cr component of 0.5, i.e., $YCbCr_{709}$ = (0.213, −0.115, 0.500). Conversion to ITU-R BT.2020 primaries results in RGB_{2020} = (0.627, 0.069, 0.016) and $YCbCr_{2020}$ = (0.213, −0.104, 0.281). Likewise, the colour with the smallest Cr component is cyan that has RGB and YCbCr values of RGB_{709} = (0, 1, 1) and $YCbCr_{709}$ = (0.787, 0.115, −0.500) respectively. Conversion to ITU-R BT.2020 will result in RGB_{2020} = (0.373, 0.931, 0.984) and $YCbCr_{2020}$ = (0.787, 0.104, −0.281). The Cr component range has therefore shrunk from [−0.5, 0.5] to [−0.281, 0.281] and in this case the constant c_{cr} is will be calculated as (0.5−(−0.5)) ÷ (0.281−(−0.281)) = 1.78.

For HEVC, if no other chroma QP offset is desired on a picture level by other means of the encoding process, the syntax elements pps_cb_qp_offset and pps_cr_qp_offset can be set equal to QPoffsetCb and QPoffsetCr, respectively. Finer control of the chroma QP offset can be achieved at the slice level.

Similarly, for AVC, if no other chroma QP offset is desired on a picture level by other means of the encoding process, the syntax elements chroma_qp_index_offset and second_chroma_qp_index_offset can be set equal to QPoffsetCb and QPoffsetCr, respectively.

An example of the effect of this method is shown in Figure 10. Figure 10 (a) shows a segment of a tone-mapped result using Formula 7-56 for an HDR/WCG image that was compressed without the use of either the luma QP or chroma QP offset modifications described above. Figure 10 (b), on the other hand, shows the same segment compressed at the same bit rate using both of these modifications. It can be seen that the large chroma artefacts, especially on the white window shutter and on the inside of the umbrella, have been ameliorated. Furthermore, the luma, especially in the wall areas, has also been improved.

(a) Without presented QP modifications (b) With presented QP modifications

Figure 10 – Image quality without (a) and with (b) the presented QP modifications

8.4 Other encoding aspects

Apart from modifying the QP allocation in the encoder, it may also be desirable for an encoder manufacturer to adjust other non-normative encoding processes in their encoders, such as the motion estimation, intra and inter mode decision, trellis quantization and rate control among others. These processes commonly consider simple distortion metrics such as mean absolute difference (MAD), or sum of squared errors (SSE), for making a variety of decisions for the decision process, and may have been tuned based on SDR content characteristics. Given, however, the earlier observations about the differences in the characteristics between SDR and HDR/WCG content, these processes may also need to be appropriately adjusted. Furthermore, other metrics may also be more appropriate for these encoding decisions. These aspects are not explored in the context of this document.

8.5 HEVC encoding

When creating the HEVC bitstream, it is recommended to set the syntax elements listed in Table 4 to the values listed in Table 4 in the sequence parameter set (SPS) of the bitstream. The syntax elements in Table 4 below are conveyed in the Video Usability Information syntax branch of the SPS defined in Annex E of the HEVC specification. They may also be duplicated and carried in various application-layer headers.

Table 4 – Recommended settings for HEVC encoding

Syntax element	Location	Recommended value
general_profile_space	profile_tier_level()	0
general_profile_idc	profile_tier_level()	2 (Main 10)
vui_parameters_present_flag	seq_parameter_set_rbsp( )	1
video_signal_type_present_flag	vui_parameters( )	1
video_full_range_flag	vui_parameters( )	0
colour_description_present_flag	vui_parameters( )	1
colour_primaries	vui_parameters( )	9
transfer_characteristics	vui_parameters( )	16
matrix_coeffs	vui_parameters( )	9
chroma_loc_info_present_flag	vui_parameters( )	1
chroma_sample_loc_type_top_field	vui_parameters( )	2
chroma_sample_loc_type_bottom_field	vui_parameters( )	2

For HDR/WCG content represented with the colour primaries of [ITU-R BT.2020] and [ITU-R BT.2100] and the PQ transfer function , the video characteristics are typically different compared to the video characteristics of SDR content represented with ITU-R BT.709 colour primaries and ITU-R BT.709 OETF (ITU-R BT.1886 EOTF) transfer function. Chroma QP adjustment, as described in clause 8.3, can be performed by adjusting and controlling the HEVC syntax elements pps_cb_qp_offset, slice_cb_qp_offset, pps_cr_qp_offset and slice_cr_qp_offset. Similarly, perceptual luma quantization, as discussed in clause 8.2, could be achieved by adjusting the syntax elements cu_qp_delta_abs and cu_qp_delta_sign_flag.

8.6 AVC encoding

When creating the AVC bitstream it is recommended to set the syntax elements listed in Table 5 to the values listed in Table 5 in the SPS of the bitstream. The syntax elements in Table 5 below are conveyed in the Video Usability Information syntax branch of the SPS defined in Annex E of the AVC specification. They may also be duplicated and carried in various application-layer headers.

Table 5 – Recommended settings for AVC encoding

Syntax element	Location	Recommended value
profile_idc	seq_parameter_set_data( )	110 (High 10)
vui_parameters_present_flag	seq_parameter_set_data( )	1
video_signal_type_present_flag	vui_parameters( )	1
video_full_range_flag	vui_parameters( )	0
colour_description_present_flag	vui_parameters( )	1
colour_primaries	vui_parameters( )	9
transfer_characteristics	vui_parameters( )	16
matrix_coefficients	vui_parameters( )	9
chroma_loc_info_present_flag	vui_parameters( )	1
chroma_sample_loc_type_top_field	vui_parameters( )	2
chroma_sample_loc_type_bottom_field	vui_parameters( )	2

For HDR/WCG content represented with the colour primaries of [ITU-R BT.2020] and the PQ transfer function, the video characteristics are typically different compared to the video characteristics of SDR content represented with ITU-R BT.709 colour primaries and ITU-R BT.709 transfer function. Chroma QP adjustment, as described in clause 8.3, can be performed by adjusting and controlling AVC syntax elements chroma_qp_index_offset and second_chroma_qp_index_offset. Similarly, perceptual luma quantization, as discussed in clause 8.2, could be achieved by adjusting the syntax element mb_qp_delta.

9 Decoding process

When the bitstream is an HEVC bitstream, the decoding process as specified in the HEVC standard is performed.

When the bitstream is an AVC bitstream, the decoding process as specification in the AVC standard is performed.

NOTE – The decoding process for HDR/WCG video is no different from the decoding process of SDR video.

10 Post-decoding processes

10.1 General

The post-decoding stage described in this document includes the following components:

a) a chroma up-conversion component that converts data from 4:2:0 to 4:4:4,

b) a conversion component that converts a fixed-point representation, i.e., 10 bits, to a floating-point representation,

c) a colour format conversion component that converts data from the non-constant luminance Y′CbCr representation back to the non-linear R′G′B′ representation, and

d) a conversion component from the non-linear data representation back to a linear data representation.

NOTE – As was also the case of the pre-encoding processing, image resolution scaling might also be desirable during this stage. For example, if the decoded data has 1920×1080 resolution, up-scaling the decoded data from 1920×1080 resolution to 3840×2160 resolution is highly likely to be performed, given the prevalence of 4K HDR/WCG displays.

Figure 11 presents a diagram of how these components could potentially be combined together to generate the desirable outcome, in a conventional manner. In this system, all blocks work independently, whereas chroma up-sampling is performed using fixed-point arithmetic and its outcome is at the same precision as the input signal.

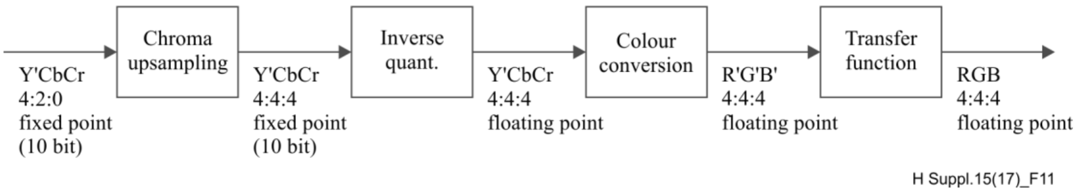

Figure 11 – Conventional post-decoding process system diagram

The various components of this stage are described in the subsequent clauses. More specifically, clause 10.2 describes the conversion process from a fixed-point representation back to a floating point representation, clause 10.3 discusses chroma up-conversion, clause 10.4 describes the colour representation conversion, i.e., from Y′CbCr back to R′G′B′ and clause 10.5 describes the conversion steps from a non-linear representation back to a linear one. Other configurations than the one depicted in Figure 2, that might not necessarily follow the same processing order and can provide different performance/complexity trade-offs, could also be used.

10.2 Conversion from a fixed-point to a floating-point representation

This process can be seen as the exact inverse of the process presented in clause 7.2.4. In particular, a fixed-point precision value can be converted to a floating-point precision value using the following formula:

$$E' = \text{Clip3}(\text{minE}, \text{maxE}, (D' - \text{offset}) \div \text{scale}) \quad (10\text{-}1)$$

The exact same values for scale and offset as in clause 7.2.4 are used according to the component type, whereas minE and maxE are equal to −0.5 and 0.5 for the chroma components respectively, and equal to 0 and 1.0 for all other colour components.

10.3 Chroma up-sampling

Chroma plane interpolation both vertically and horizontally is performed to convert the 4:2:0 NCL Y′CbCr signal to a 4:4:4 representation. Similar to the down-conversion process in clause 7.2.3, this step needs to again account for the siting of the chroma components compared to those of the luma (Figure 4).

It is quite likely that FIR linear filters would be used by many implementations for this process. Similar to the down-conversion case, it has been reported that non-linear or adaptive filters could result in improved objective performance. The simple two-phase resampling filter shown in Table 6 was used for the experiments conducted for the preparation of this report. The same filter is applied both vertically and horizontally.

NOTE – This is essentially a "Lanczos 2" filter. Higher-precision and higher-order filters could potentially be used, especially when up-sampling content of very high quality or with no compression.

Table 6 – Two-phase chroma resampling filter

Phase p	Interpolation filter coefficients			
	f_C[p, −1]	f_C[p, 0]	f_C[p, 1]	f_C[p, 2]
0	0	1	0	0
1	−1/16	9/16	9/16	−1/16

10.4 Colour representation conversion: non-constant luminance Y′CbCr to R′G′B ′

Conversion from the non-constant luminance Y′CbCr representation back to the R′G′B′ representation can be performed using the following formula:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \mathbf{A} \begin{bmatrix} Y' \\ Cb \\ Cr \end{bmatrix} = \mathbf{W}^{-1} \begin{bmatrix} Y' \\ Cb \\ Cr \end{bmatrix} = \frac{1}{\det(\mathbf{W})} \operatorname{adj}(\mathbf{W}) \begin{bmatrix} Y' \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} a_{RY} & a_{RCb} & a_{RCr} \\ a_{GY} & a_{GCb} & a_{GCr} \\ a_{BY} & a_{BCb} & a_{BCr} \end{bmatrix} \begin{bmatrix} Y' \\ Cb \\ Cr \end{bmatrix} \qquad (10\text{-}2)$$

Since $\mathbf{A} = \mathbf{W}^{-1}$, all the matrix coefficients in **A** can be calculated directly from w_{YR}, w_{YG}, w_{YB}. In particular, given the characteristics and precision of the coefficients used in clause 7.2.2, it is possible to compute that:

$$a_{RY} = a_{GY} = a_{BY} = 1 \qquad (10\text{-}3)$$

$$a_{RCb} = a_{BCr} = 0 \qquad (10\text{-}4)$$

$$a_{RCr} = 2 * (1 - w_{YR}) \;=\; 1.474\ 6 \qquad (10\text{-}5)$$

$$a_{GCb} = -\frac{2 * w_{YB} * (1 - w_{YB})}{w_{YG}} = -0.164\ 553\ 126\ 843\ 660 \qquad (10\text{-}6)$$

$$a_{GCr} = -\frac{2 * w_{YR} * (1 - w_{YR})}{w_{YG}} = -0.571\ 353\ 126\ 843\ 660 \qquad (10\text{-}7)$$

$$a_{BCr} = 2 * (1 - w_{YB}) \;=\; 1.881\ 4 \qquad (10\text{-}8)$$

If high precision is possible, it is recommended that the following matrix is used:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.000\ 000\ 000\ 000\ 00 & 1.474\ 600\ 000\ 000\ 00 \\ 1 & -0.164\ 553\ 126\ 843\ 66 & -0.571\ 353\ 126\ 843\ 66 \\ 1 & 1.881\ 400\ 000\ 000\ 00 & 0.000\ 000\ 000\ 000\ 00 \end{bmatrix} \qquad (10\text{-}9)$$

For systems with limited precision, a lower precision representation of the above matrix, e.g., retaining only 6 digits of precision, could be used instead.

10.5 Conversion from a non-linear to a linear light representation: R′G′B′ to RGB

Conversion from a non-linear to a linear light representation is performed using an electrical-optical transfer function (EOTF). In this document, the exact inverse of the process described in clause 7.2.1 is used.

More specifically, the linear light intensity signal L_o can be computed from the non-linear representation V, which takes values in the range [0, 1], as follows:

$$L_o = \mathrm{EOTF}(V) = \left(\frac{\mathrm{Max}\left(\left(V^{\frac{1}{m}} - c_1 \right), 0 \right)}{c_2 - c_3 * V^{\frac{1}{m}}} \right)^{\frac{1}{n}} \qquad (10\text{-}10)$$

where c_1, c_2, c_3, m and n were defined in clause 7.2.1. In Figure 12, notable points on the plot include: non-linear real-value signal value 0.5 (corresponding to 10-bit narrow code level 502) maps to 0.01 real-value signal output (0.01 * 10 000 = 100 cd/m^2), while input value 0.9 (10-bit code level 853) maps to 0.4 (4 000 cd/m^2).

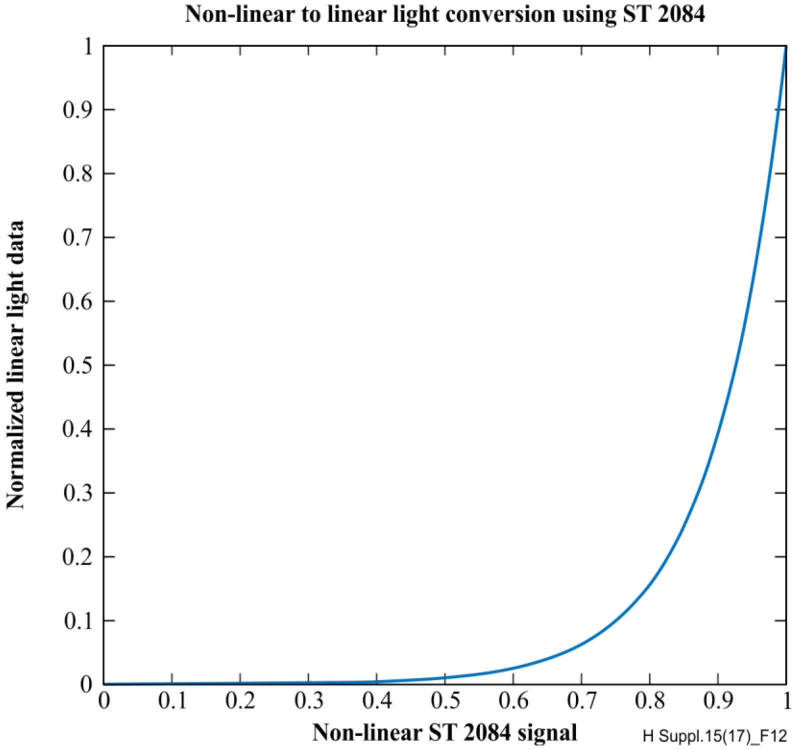

Figure 12 – Behaviour of the PQ EOTF

This process is applied to all R′, G′ and B′ non-linear representations, resulting in their linear light counterparts R, G and B as follows.

$$R = EOTF(R') \quad (10\text{-}11)$$

$$G = EOTF(G') \quad (10\text{-}12)$$

$$B = EOTF(B') \quad (10\text{-}13)$$

LUTs, as was also the case for the linear to non-linear conversion, could also be used instead of using Formula 10-10 directly. In this scenario, a non-uniform LUT interpolator is also recommended.

Appendix I

Supplemental enhancement information (SEI) messages

This appendix provides short descriptions of SEI messages specified in AVC or HEVC that can be particularly relevant for HDR/WCG video.

I.1 Mastering display colour volume SEI message

The mastering display colour volume SEI message identifies the colour volume (the colour primaries, white point and luminance range) of a display considered to be the mastering display for the associated video content, e.g., the colour volume of a display that was used for viewing while authoring the video content. If mastering display colour volume information is included, it is recommended that mastering display colour volume SEI messages are included at least at each random access point access unit (RAPAU). The information provided in the mastering display colour volume information SEI message is recommended to apply until, but not necessarily including, the next RAPAU. If multiple mastering display colour volume SEI messages are included in the bitstream between the start of two RAPAUs, then those SEI messages are recommended to have the same content. Table I-1 shows an example of what values the mastering display colour volume SEI message would contain in the case that the mastering display used P3 colour primaries, as per [SMPTE RP 431-2], the D65 white point and a luminance range of 0.000 1 cd/m^2 to 2 000 cd/m^2, inclusive.

Table I.1 – An example mastering display colour volume SEI message

Syntax element	Example value
display_primaries_x[0]	13 250
display_primaries_y[0]	34 500
display_primaries_x[1]	7 500
display_primaries_y[1]	3 000
display_primaries_x[2]	34 000
display_primaries_y[2]	16 000
white_point_x	15 635
white_point_y	16 450
max_display_mastering_luminance	20 000 000
min_display_mastering_luminance	1

If the decoder contains an interface for the output of mastering display colour volume information and the bitstream contains mastering display colour volume SEI messages it is recommended that the decoder outputs mastering display colour volume information synchronously to the first picture for which the SEI message apply.

I.2 Content light level information SEI message

The content light level information identifies upper bounds for the nominal target brightness light level of the coded pictures. [Blu-ray 2015] provides guidelines that the value of MaxFALL should not exceed the value of 400 cd/m^2 and that only a small percentage of the picture, such as specular highlights, can exceed 1 000 cd/m^2. For the BD-ROM 3.0 application, these measurements are computed off-line during a complete pass through the content and before the final values for MaxFALL and MaxCLL are computed. Live or broadcast services may set these values to pre-determined levels without any actual measurement on the content.

I.3 Ambient viewing environment SEI message

The ambient viewing environment SEI message describes the ambient viewing environment (the HRVE) that was targeted for viewing when mastering the associated video content. It conveys the environmental illuminance and chromaticity coordinates of the mastering nominal ambient viewing environment.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	**Audiovisual and multimedia systems**
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2017