 ISO/IEC JTC 1/SC 29/AG 3 N0034

**ISO/IEC JTC 1/SC 29/AG 3**

**MPEG Systems   
Convenorship: KATS (Korea, Republic of)**

**Document type:** Output Document

**Title:** Draft of White paper on Essential Video Coding (EVC)

**Status:** Approved

**Date of document:** 2021-07-27

**Source:** ISO/IEC JTC 1/SC 29/AG 3

# Expected action: None

# Action due date: None

**No. of pages:** 11 (without cover page)

**Email of Convenor:** kyuheonkim@khu.ac.kr

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29ag3>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/AG 3**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC 1/SC 29/AG 3 N** **00034**

**Online – July 2021**

|  |  |
| --- | --- |
| **Source:** | **Convenor of ISO/IEC JTC 1/SC 29/AG 03** |
| **Status:** | **Approved by AG 03** |
| **Subject:** | **Draft of White paper on Essential video coding (EVC)** |
| **Date:** | **July 27, 2021** |
| **Serial Number:** | **20735** |

**Abstract**

This document provides a draft text for EVC white paper.

1. **Introduction**

The MPEG-5 Essential Video Coding (EVC) standard was finalized by the ISO/IEC Moving Picture Experts Group (MPEG) in April 2020 [1]. The main objective of the EVC standard development was to provide a significantly improved compression efficiency over existing video coding standards with timely publication of licensing terms.

This white paper reviews the architecture and building blocks of EVC, which were carefully selected considering compression capability, complexity, and timely publication of licensing to achieve the goal of the MPEG-5 part1 project.

1. **Background**

There is a constant demand for more efficient video coding technologies, however coding efficiency is not the only factor which determines the industry choice of video coding technology for products and services. Video coding technologies should address the needs of existing and emerging real-world use cases and should also be easy to adopt from both technological and business perspectives.

As motivated by business perspectives but still existing demands on a high performance license-friendly video codec, MPEG issued a requirement to develop a new video coding standard that addresses both technical and business requirements that are not adequately met satisfied by previous video coding standards [2].

To achieve the requirements, the EVC standard was developed with the royalty-free based Baseline profile as its base and a royalty bearing Main profile having a small number of improved coding tools on top of the Baseline profile.

The Baseline profile contains technologies assessed to be over 20 years to achieve a royalty free codec. The Baseline profile builds a video codec using only conventional coding technologies which consist of traditional methods from the early 1980s to the end of the 1990s. On top of the Baseline profile, a small number of tools were added to improve coding performance for the Main profile. Each of these tools in the Main profile provide a significant improvement in terms of coding efficiency and are capable of being cleanly switched off on an individual basis.

The purpose of such a licensing-friendly video codec is to reduce business uncertainty by minimizing the risk of excessive delay in understanding the full licensing cost of launching products or services based on the standard.

1. **Key technology and feature**

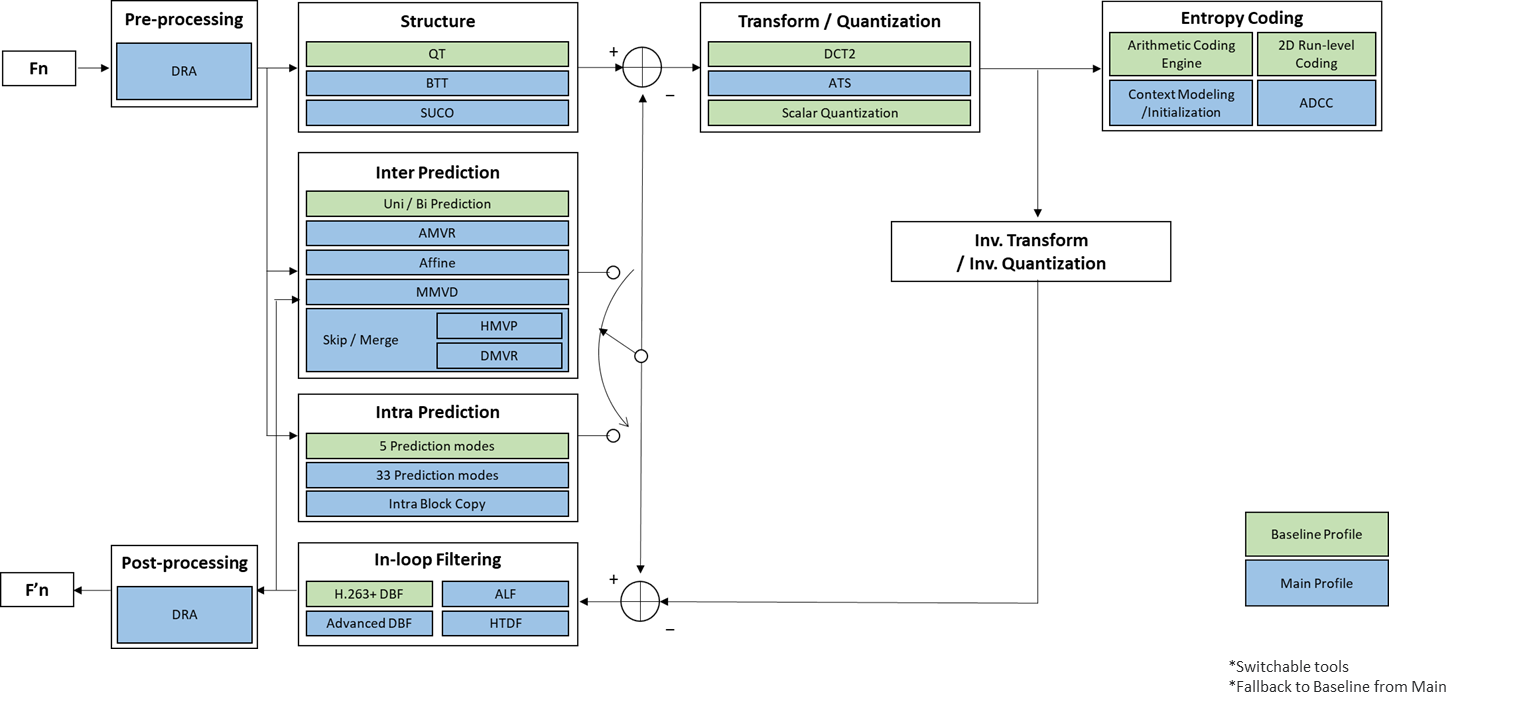


Figure 1. Block diagram of EVC coding tools

Figure 1 shows a tool-level summary of EVC, which is illustrated in a block diagram of tools associated with its profile. All of tools in Baseline profile were made based on technologies assessed as available for over 20 years. The tool set of Main profile added on top of the basic set of Baseline profile. The tools for Main profile were carefully chosen by considering an interaction between the tools to enhance a performance with the limited number of coding tools.

* 1. **Tools of Baseline profile**
     1. **Coding structure**

Coding structure would be the most effective tool to dramatically improve coding gain. Historically, new video codecs have always been proposed with new coding structures. A 16x16 macroblock structure was employed until AVC. Since coding structure based on 16x16 blocks gave poor coding performance in the case of large resolutions such as 1080p or 4K, HEVC and subsequent video coding standard developments employed larger block structures such as 64x64 with a flexible splitting mechanism to divide a large coding block into small blocks. The coding structure in baseline profile employed a quad-tree based coding structure which can use blocks up to 64x64. The quad-tree based coding algorithm was introduced in the early 1990s and an efficient splitting method was disclosed in 1994 [3].

* + 1. **Intra prediction**

Intra prediction is the technology to exploit spatial correlation between neighbouring pixels. While prediction is performed in the transform domain in H.263 and MPEG-4 visual, AVC introduced prediction in the pixel domain by referencing the neighbouring pixels of the previously coded blocks [4][5]. Higher coding gain has been achieved by increasing the number of prediction directions. In Baseline profile, 5 directional prediction modes are employed; DC, horizontal (H), vertical (V), diagonal left (DL), diagonal right (DR). A codeword for prediction mode of the current block is adaptively assigned by using a mapping table between symbol and prediction mode which is selected based on the prediction modes of neighbouring upper and left blocks.

* + 1. **Inter prediction**

Baseline profile exploits three neighboring motion vectors and a motion vector of temporally co-located blocks. After choosing one of the candidate motion vectors as a predictor, the index of the predictor is coded. The difference between the motion vector for the current block and the predictor is not necessarily coded. If the predictor is good enough, the motion vector difference and a block residue are not coded, which is called the skip mode. The bi-directional prediction is a linear combination of two motion compensated blocks that involve two motion vectors, a forward and backward motion vector.

* + 1. **Transform and Quantization**

As numerous codecs for video compression have generally used, Baseline profile also adopts a discrete cosine transform (DCT), which is well-known as having good energy compaction efficiency. Transforms are applied to a residual block between an original block and the corresponding prediction block, as a conventional hybrid video codec does. Since transforms are applied to coding blocks, the transform size is equal to the coding block size, i.e. from 2x2 to 64x64. After the transform is conducted, scalar quantization is applied to the transformed coefficients. The range of the quantization parameter (QP) is from 0 to 51 and a scaling factor (SF) corresponding to each QP is defined by a look-up table.

* + 1. **Loop filter**

Most block-based video coding schemes introduce blocking artifacts, which are very noticeable at low bitrates. The goal of a loop filter is to reduce the blocking artifacts and improve visual quality. To remove blocking artifacts, post filtering and loop filtering have been used. The loop filters, unlike the post filters, operate inside of coding loop and they improve the quality of reference frames. Since refining reference frames results in improvement in visual quality and coding efficiency, loop filtering plays an important role in video coding. In this Baseline profile, a loop filter based on H.263 Annex J [4] was employed to increase objective and subjective image quality in this section. The algorithm is simple enough to operate in the coding loop, but is able to successfully remove blocking artifacts using the quantization parameter.

* + 1. **Entropy coding**

For entropy coding in video and image compressions, Huffman and arithmetic coding have been employed. The state of the art entropy coding scheme is context-based adaptive binary arithmetic coding (CABAC), introduced in the AVC standard [6]. Instead of CABAC, the binary arithmetic coding scheme in JPEG Annex D is applied as the entropy coding engine of the Baseline profile [7]. After a binarization process of the given symbol, an arithmetic coding engine encodes each binary value with the corresponding context that stores the occurrence probability of a given value. After each binary value is encoded, the probability is updated by using a look-up table and the binary value of symbol is stored in the corresponding context. To code the transformed and quantized coefficient values, run/level symbols are generated after scanning with a zig-zag pattern. Each run or level symbol is binarized by unary coding and the binary value is coded with the corresponding run or level contexts. The sign value and the last symbol indicate whether the level is the last one in the block should be followed to each level. The sign value is coded with fixed length coding and the last symbol is coded with the arithmetic engine.

* 1. **Tools of Main profile**
     1. **Coding structure**

Main profile supports of a flexible block partitioning structure, which is based on the binary and ternary trees mixture scheme (BTT) with Split Unit Coding Order (SUCO), for efficient and flexible representation of video content with various resolutions. BTT has CU shapes described by the ratio between the width and height of a block as shown in Figure 2. For instance, if the width and height of a block are the same it can be represented as a 1:1 ratio CU or a square CU, and if the width is equal to 64 and the height is equal to 16 it can be represented as a 1:4 ratio CU. CU partitioning is conducted based on the allowed CU shapes and their allowed maximum and minimum sizes.

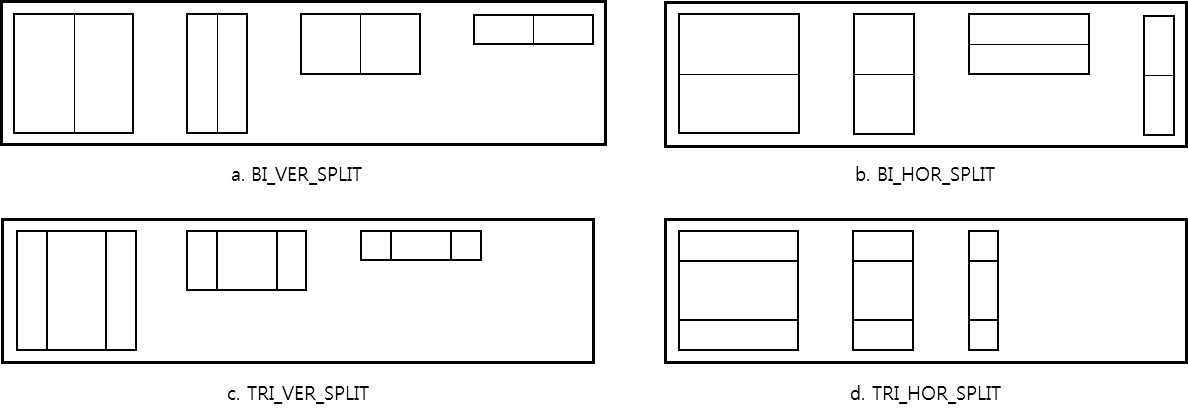
****

Figure 2. Binary / Ternary split modes

As shown in Figure 3, the Split unit coding order (SUCO) enables a more flexible coding order, such as left to right (L2R) and right to left (R2L), to allow intra prediction from right reference pixels and inter prediction with right motion vector predictors. If an SU is partitioned horizontally (horizontal splitting), a flag is signalled to indicate L2R or R2L coding order of partitioned units. If an SU is partitioned by a quad tree structure, a flag is shared for the two above units and two bottom units. If no flag is signalled for coding order of an SU, the coding order follows its parent’s SU coding order.



Figure 3. Allowed coding order in Quad-tree / Binary / Ternary split modes

* + 1. **Intra prediction**

To exploit spatial correlation efficiently based on flexible coding structure, a total of 33 intra prediction modes for luma component and 5 modes for chroma component are applied. DC, Bi-linear, Planer, and DM modes are similar to that of AVS2, with straightforward extension for flexible block size while plane and angular prediction mode are different from that in AVS2 [8].

* + 1. **Merge with Motion Vector Difference**

MMVD provides a new motion vector expression method with simplified signaling. Similar to the skip and merge modes in HEVC, MMVD makes a candidate list from neighboring motion information but MMVD can cover more extended motions that are not limited to the neighboring motions. In order to construct more precise candidates, a starting point, a motion magnitude, and a motion direction are utilized. The use of MMVD eliminates a lot of MVD signaling; instead, signals on the related indices are delivered to the decoder.

* + 1. **Adaptive motion vector resolution**

Adaptive motion vector resolution (AMVR) supports multiple motion vector resolutions. Supported motion vector resolutions in Main profile range from 1/4–pel to 4–pel (1/4–pel, 1/2–pel, 1–pel, 2–pel and 4–pel). Information about the motion vector resolution is signalled at the CU level when MVD information is signalled. Depending on the resolution of CU, both motion vector (MV) and motion vector predictor (MVP) of the CU are adjusted.

* + 1. **Affine Mode**

Affine prediction of the Main profile allows the use of three affine motion modes: namely four and six parameters model modes as well as an affine merge mode. In EVC, the affine motion field for a CU is described by MVs of two control points located at the top-left and top-right corners (four parameter model) or MVs of three control points located at the top-left, top-right and bottom-left corners (six parameter model). In four and six parameters model modes, the control point MVs for the current CU are signaled in the bitstream. For affine merge mode, the control point MVs of the current CU are derived based on the motion information of the neighboring CUs. When merge or skip mode is applied, and both width and height for the CU are larger than or equal to 8, an affine flag in CU level is signaled in the bitstream to indicate whether affine merge mode is used. In this mode, the merge candidate index with a maximum value of 4 is signaled for specifying which motion information candidate in the affine merge candidate list is used for the CU.

* + 1. **Decoder-side motion vector refinement**

DMVR method operates with the two motion vectors of the bi-prediction which are further refined by a bilateral matching process. In bi-prediction mode, refined MVs are searched around the initial MVs in the reference picture list L0 and reference picture list L1. The DMVR searching process consists of an integer sample MV offset search and fractional sample MV refinement. The integer sample MV searching calculates the distortion between the two candidate reference blocks in the reference picture list L0 and list L1. The SAD between the reference blocks based on each MV candidate around the initial MV is calculated. The MV candidate with the lowest SAD becomes the refined MV and used to generate the bi-predicted signal. The search range is 2 integer luma samples from the initial MVs. The fractional sample refinement is conditionally invoked based on the output of the integer sample search stage. To save the calculational complexity, the fractional sample refinement is derived by using a parametric error surface equation, instead of additional search with SAD comparison. In parametric error surface based sub-pixel offsets estimation, the center position cost and the costs at four neighboring positions from the center are used to fit a 2-D parabolic error surface equation.

* + 1. **History-based Motion Vector Prediction**

History-based motion vector prediction (HMVP) method is an inter coding tool which can be applied to both merge candidate list and Advanced motion vector prediction (AMVP) candidate list construction process. In HMVP, a table of HMVP candidates is maintained and updated on-the-fly. Whenever a non-affine inter coded block is decoded, the decoded motion information is used to update the HMVP table in the last position following a first-in-first-out (FIFO) rule to remove and add entry. The HMVP table size is set to be 23 and subsampling fetching process is employed to avoid pruning over the HMVP table and thus reduce complexity. For each of the HMVP entry, a single MV and refIdx for uni-prediction or two MVs and two refIdx for bi-prediction are stored.

* + 1. **Adaptive Loop Filter**

To suppress coding artifacts, improve visual and objective quality of the decoded and reference pictures, decoded samples are filtered with ALF. Parameters of ALF are signaled in independent NAL units, called Adaptation Parameter Set (APS) which can be referred from slice or tile coding unit. For luma filtering, two types of diamond filter patterns (5x5 and 7x7) are defined and for chroma samples, only 5x5 pattern is used. Each ALF filter structure may consist of up to 25 different luma filters and utilized luma filter is selected through a classification process for each 4x4 block. For chroma filtering, a single filter per APS can be signaled and used in slice or tiles group.

To reduce bit overhead, several filter coefficient prediction techniques are employed. To benefit from symmetrical properties of filters, utilized ALF employs a filter coefficient transformation process, thus reducing the number of independent filter coefficients to 13. Following this, a set of static filter coefficients is provided to the decoder as side information and used as predictors for signaled ALF coefficients.

And finally, to benefit from temporal correlation, employed ALF design utilizes re-usage of the ALF coefficients signaled earlier. Each signalled APS consisting ALF is identified by the unique adaptation parameter set ID which is used for referencing the relevant ALF information from other syntax elements, e.g. from slice or tiles group. All signaled APS with unique set id value are stored in APS buffer size up to 32 entries. To enable the RA coding configuration, encoder’s choice of the APS ID usage is constrained. For example, to keep the temporal scalability, only APS from the same or lower temporal layers can be referred

* + 1. **Hadamard transform domain filter**

In addition to deblocking filter and ALF, Hadamard transform domain filter aiming to reduce ringing artifacts caused by quantization of residual coefficients is introduced in EVC. Hadamard transform domain filter (HTDF) is applied to luma reconstructed blocks when the quantization parameter (QP) is larger than 17. The transform core is 2x2 Hadamard transform, which results in that HDTF is 3x3 low pass smoothing filter. The filter parameters are explicitly derived from the coded information, including transform block size and QP. The order of filtering process in EVC is HDTM, deblocking and ALF.

* + 1. **Adaptive transform selection**

Adaptive transform selection (ATS) is exploited in the Main profile. In addition to traditional DCT II transform, DST-VII and DCT-VIII can be applied for intra and inter predicted residual. For intra coded block, a flag is used to signal to the decoder whether ATS applied or not. Basically, the encoder decides the use of ATS based on the RDO process at CU level. If encoder selects usage of ATS in a CU as core transform, two more flags are signaled to the decoder to indicate which type is used, for the horizontal and vertical directions respectively. For an inter-predicted CU that contains residuals, a flag is signaled to indicate whether the whole residual block or a sub-part of the residual block should be decoded. When only a sub-part of the residual block is coded, the part of the residual block is coded with inferred transform type (DST-VII or DCT-VIII) and the other part of the residual block is zeroed out.

* + 1. **Advanced coefficient coding**

Transform coefficients of the coded block (residual data) after quantization are scanned in a predefined scan pattern and entropy coded. To employ statistical properties of transform coefficients, advanced coefficient coding (ADCC) utilizes the bit-plane like coding approach. In particular, transformed coefficients are scanned and coordinates (X and Y) of the last non-zero transform coefficient in the scan order are signaled. Following this, coefficients are parsed in inverse zig-zag scan order and processed in chunks of the maximal size of 16. Coefficients within each processing chunk are signaled as a sequence of significance (i.e., sigMapFlag) and levels flags (i.e., flagLevelA and flagLevelB), sign flag and remaining levels (i.e., levelRem). Among these symbols, the bins of sigMapFlag, flagLevelA and flagLevelB are encoded with adaptive context models; signFlag and binarized bins of levelRem are encoded through by-pass mode. To decrease number of context-coded bins, the explicit flagLevelA and flagLevelB are adaptively switched into levelRem coding, which is binarized with golomb code and encoded with bypass mode with equal probability.

1. **Profiles and levels**

EVC specified four profiles including “Baseline profile”, “Main profile”, “Baseline Still Picture profile”, and “Main Still Picture profile”.

Similar to previous video compression standards, EVC defines conformance points in terms of profiles (combinations of decoding tools and associated bitstream syntax that is expected to be interpreted by a decoder), and levels (typically maximum sizes of pictures and frame rates, maximum bit rate, buffer capacity etc.). Currently, a total of 13 levels, ranging in support for typical picture sizes as small as VGA at the low end, up to as large as 8Kx4K ultra HD at the high end, are specified.

1. **Performance**

The MPEG carried out a formal subjective verification test of EVC Main and Baseline profiles for standard dynamic range (SDR) content and high dynamic range (HDR) and wide colour gamut (WCG) content, respectively [9][10]. The purpose of the verification test was to confirm that the coding efficiency objective for the EVC standard has been met: achieving a substantial bit-rate reduction at the same level of subjective visual quality relative to HEVC Main profile for EVC Main profile and AVC High10 profile for EVC Baseline profile, respectively.

* 1. **Results on SDR sequence**

The assessment included SDR ultra high definition (UHD, a.k.a. 4K, 3840×2160) test sequences encoded in random access (RA) configuration and SDR full high definition (HD, a.k.a. 2K, 1920×1080) test sequences encoded in low delay (LD) configuration.

|  |  |
| --- | --- |
|  |  |
| 1. CatRobot | 1. DrivingPOV3 |
|  |  |
| 1. Marathon2 | 1. BarScene |

Figure 4. Results on SDR UHD resolution with RA configuration

|  |  |
| --- | --- |
|  |  |
| 1. BarScene | 1. DrivingPOV |
|  |  |
| 1. Metro | 1. RushHour |

Figure 5. Results on HD resolution with LD configuration

The average bit rate saving of the EVC profiles relative to the anchors were computed from the MOS vs. bit rate data for each sequence, in the same manner that was done in [11], to further quantify the bit rate savings achieved.

The bit rate savings were averaged over the whole range where the same MOS scores for EVC Main profile, EVC Baseline profile, HEVC and AVC could be interpolated from subjective test results shown in Figure 4 and 5.

The MOS BD-rate calculation [12] shows that the average bit rate saving for EVC Main profile compared to HEVC Main10 profile was approximately 39% using UHD SDR content encoded in random access configuration, and approximately 41% using HD SDR content encoded in low delay configuration. The average bit rate saving for EVC Baseline profile compared to AVC High10 profile was approximately 39% using UHD SDR content encoded in random access configuration, and approximately 34% using HD SDR content encoded in low delay configuration.

* 1. **Results on HDR/WCG sequence**

Subjective testing in HDR category was conducted through a subjective evaluation, comparing coding performance of the EVC Main profile to the HEVC Main 10 profile for video sequences in BT.2100/PQ representation [13].

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) Chimera5 | (b) Chimera7 | (c) Meridian2 |

Figure 6. Results on UHD resolution with RA configuration

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1. Chimera5 | 1. Chimera7 | 1. Hurdles |

Figure 7. Results on HD resolution with RA configuration

Figure 7 shows the visual testing results in HDR/WCG category. The average bit rate savings for EVC Main profile with HDR-UHD and HDR-HD are estimated at approximately 36% and 35% respectively compared to HEVC Main10 profile.

1. **Resources**

EVC related documents are available on https://dms.mpeg.expert/, and the ETM reference software repository is available at http://mpegx.int-evry.fr/software/MPEG/Video/EVC/ETM. The open-source encoder and decoder of EVC, which are XEVE [13] and XEVD [14], was released at public repository, respectively.

1. **Reference**

[1] ISO/IEC 23094-1:2020 Information technology — General video coding — Part 1: Essential video coding.

[2] "Requirements for a New Video Coding Standard", ISO/IEC JTC1/SC29/WG11 N17928, Oct. 2018.

[3] G.J. Sullivan and R.L. Baker, “Efficient quadtree coding of images and video”, IEEE Trans-actions on Image Processing, 327-331, May 1994.

[4] Video Coding for Low bit rate Communication, ITU-T Recommendation H.263, version 1, Nov. 1995; version 2, Jan. 1998; version 3, Nov. 2000.

[5] Coding of audio-visual objects—Part 2: Visual”, ISO/IEC 14496-2 MPEG-4 visual, Apr. 1999.

[6] "Advanced Video Coding (AVC)”, ITU-T Recommendation H.264 and ISO/IEC 14496—10, May 2003.

[7] Information technology-digital compression and coding of continuous-tone still images-requirements and guidelines. International Telecommunication Union. CCITT recommendation, 81, p.09.

[8] He, Zhichu, Lu Yu, Xiaozhen Zheng, Siwei Ma, and Yun He. “Framework of AVS2-video coding”, IEEE International Conference on Image Processing, 1515-1519, 2013.

[9] "Report on Essential Video Coding Compression Performance Verification Testing for HDR/WCG Content", ISO/IEC JTC 1/SC 29/WG 04 N0030, Oct. 2020.

[10] "Report on Essential Video Coding compression performance verification testing for SDR Content", ISO/IEC JTC 1/SC 29/WG 04 N0047, Jan. 2021.

[11] T .Tan, M. Mrak, V. Baronicini, N. Ramzan, “Report on HEVC compression performance verification testing”, JCTVC-Q1011, April 2014.

[12] Gisle Bjøntegaard, “Improvements of the BD-PSNR model”, ITU-T SG16/Q6, 35th VCEG Meeting, Berlin, Germany, July, 2008, Doc.VCEG-AI11.

[13] International Telecommunication Union – Radio Communication Sector, Recommendation ITU-R BT.2100-2: Image parameter values for high dynamic range television for use in production and international programme exchange.

[14] eXtra-fast Essential Video Encoder (XEVE). [Online]. Available: https://github.com/mpeg5/xeve

[15] eXtra-fast Essential Video Decoder (XEVD), [Online]. Available: https://github.com/mpeg5/xevd