|  |
| --- |
| **INTERNATIONAL ORGANIZATION FOR STANDARDIZATION ORGANISATION INTERNATIONALE DE NORMALISATION ISO/IEC JTC 1/SC 29/WG 5 MPEG JOINT VIDEO CODING TEAM WITH ITU-T SG 16** |
| **ISO/IEC JTC 1 / SC 29 / WG 5 N 177** |
| **Teleconference, 11–20 January 2023** |
| |  |  | | --- | --- | | **Title:** | **Preliminary WD of TR: Optimization of encoders and receiving systems for machine analysis of coded video content** | | **Source:** | **Convenor (Jens-Rainer Ohm)** | | **Type:** | **Project** | | **Subtype:** | **Draft (Other)** | | **Status:** | **Approved** | | **Date:** | **2023-04-20** | | **Expected Action:** | **Info** | | **Action due date:** | **N/A** | | **No. of pages** | **7** (without this cover page) | | **Email of convenor:** | **ohm @ ient . rwth-aachen . de** | | **Committee URL:** | **https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-5** | |

|  |  |
| --- | --- |
| **Joint Video Experts Team (JVET)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29**  29th Meeting, by teleconference, 11–20 January 2023 | Document: JVET-AC2030-v1 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **Optimization of encoders and receiving systems for machine analysis of coded video content (draft 1)** | | |
| *Status:* | Output document of JVET | | |
| *Purpose:* | Report | | |
| *Author(s) or Contact(s):* | **Jie Chen** (Alibaba)  **Christopher Hollmann** (Ericsson)  **Shan Liu** (Tencent) | Email: | jiechen.cj@alibaba-inc.com  christopher.hollmann@ericsson.com  shanl@tencent.com |
| *Source:* | Editors | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

This document provides a description of optimization techniques for video encoders and receiving systems for machine consumption. It provides a concept-level overview of recent practices and references to technical papers that describe further details. It provides comments on some technical aspects and identifies situations where caution must be taken when interpreting the results.

# Scope

This document provides summary information about optimizations for encoders and receiving systems for conducting machine tasks on coded video content using video coding standards developed in the ITU-T SG 16 Visual Coding Experts Group (VCEG) and ISO/IEC JTC 1/SC 29 Moving Picture Experts Group (MPEG) communities. This document does not represent an authoritative recommendation for how encoders and receiving systems should be optimized. It merely describes technologies that have recently been studied and demonstrated benefits to improvement on coding efficiency for machine tasks.

# Normative references

There are no normative references in this document.

# Terms and definitions

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

* ISO Online browsing platform: available at <https://www.iso.org/obp>
* IEC Electropedia: available at <http://www.electropedia.org/>

## Definitions

**Machine consumption**

A machine vision task such as object detection, instance segmentation or object tracking etc. being used on the decoded video.

## Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

|  |  |
| --- | --- |
| AVC | Advanced Video Coding (Rec. ITU-T H.264 | ISO/IEC 14496-10) |
| BD-rate | Bjøntegaard Delta bit rate |
| CfE | Call for Evidence |
| CfP | Call for Proposals |
| HEVC | High Efficiency Video Coding (Rec. ITU-T H.265 | ISO/IEC 23008-2) |
| JVET | Joint Video Experts Team (for development of VVC) |
| mAP | Mean Average Precision |
| MOTA | Multiple Object Tracking Accuracy |
| MPEG | Moving Picture Experts Group |
| PSNR | Peak Signal-to-Noise Ratio |
| QP | Quantization Parameter |
| RoI | Region of Interest |
| TID | Temporal Identifier |
| VCEG | Video Coding Experts Group |
| VVC | Versatile Video Coding (Rec. ITU-T H.266 | ISO/IEC 23090-3) |
| *Y′CBCR* | Colour space representation commonly used for video/image distribution, also written as *YUV* |
| *YUV* | Colour space representation commonly used for video/image distribution, also written as *Y′CBCR* |

# Overview

Similar to most video processing systems, there are four main processing steps, as shown in Figure 1. As the scope of this report focuses on optimizations for encoders and receiving systems, technologies for pre-processing, encoding and post-processing are described in this report. The decoding process, on the other hand, is fully specified in the respective Versatile Video Coding (VVC), High Efficiency Video Coding (HEVC) and Advanced Video Coding (AVC) video coding standards, amongst others, and a decoder must fully comply to the intended profile and level to properly decode and output the reconstructed video samples from a given input bitstream.

Input video

Pre-processing

Encoding

Decoding

Post-processing

Machine consumption

Figure 1. General processing pipeline

A description of different pre-processing methods can be found in clause 6. Encoder optimizations and processes are detailed in clause 7 and post-processing methods are described in clause 8. An overview of the commonly used practices for evaluating these kinds of optimizations in JVET and MPEG communities can be found in clause 5.

In July 2019 the Requirements subgroup of ISO/IEC JTC 1/SC 29/WG 11 (MPEG) started exploration activities in the area of video coding for machines, with the initial goal of defining use cases, requirements and evaluation framework. This work, which was continued by ISO/IEC JTC 1/SC 29/WG 2, resulted in a formal call for evidence (CfE) in January 2021 [5]. As the responses to this CfE showed evidence of improvements over existing compression standards, a call for proposals (CfP) was issued in April 2022 [6].

Some responses to the CfP contained elements that can be used without modifications to existing video compression standards, and one response produced bit streams that were compliant to the VVC with pure encoder optimization [3]. As encoder optimization methods demonstrated significant compression benefits to machine consumption, compared with the anchor configuration that is setup for human consumption, JVET decided to conduct a study on optimization technologies for encoders and receiving systems for machine consumption. The outcome of this study is documented in this report.

# Evaluation methodology

A set of assessment metrics and measurement points are used for the evaluation of encoder optimization technologies for machine consumption. An overview evaluation framework is shown in Figure 2. Here the input video is encoded to generate a bit stream. This bit stream is then decoded, and the decoded video is used for the machine consumption task. Note that in this diagram the “Encoder” includes both pre-processing and encoding steps as in Figure 1, and the “Decoder” includes both decoding and post-processing steps shown in Figure 1.

Task performance calculation

Input video

Encoder

Decoder

Output video

bit rate

Ground truth

Machine consumption

YUV PSNR calculation

Figure 2. Measurement points used to evaluate technologies

The bit rate is determined based on the encoded bit stream and parameters of the input video such as frame rate and the number of total frames. The following equation can be applied:

(1)

Encoding for video distribution is ordinarily performed in the *Y′CBCR* domain (nicknamed *YUV* herein for brevity and ease of typing). For standard-dynamic range video, the distortion metric primarily used in the video coding standardization community has been the Peak Signal to Noise Ratio (PSNR). A detailed description on how to determine PSNR can for example be found in [2].

For different machine consumption tasks different metrics are used. For example, the performance for the object detection and instance segmentation tasks is usually measured in mean average precision (mAP). This metric indicates what percentage of objects are correctly identified by having sufficient overlap between the detected object and the ground truth as well as being assigned the correct object class. Then the share of correctly identified objects for each class is determined, and finally the score for each class is averaged. There are different variants of this metric:

* mAP@0.5: For an object to be counted as correctly identified the overlap between the detected bounding box and the bounding box in the ground truth needs to be at least 0.5. Sometimes this variant of the mAP metric is also referred to as mAP50.
* mAP@[0.5:0.05:0.95]: In this variant a total of ten mAP scores with increasing overlap thresholds are calculated. For each iteration the overlap threshold is increased by 0.05, starting at 0.5 and reaching all the way to 0.95. Once all ten scores are determined, the average of these scores is calculated to produce the final mAP performance.

(2)

For the object tracking task, performance is often measured in multiple object tracking accuracy (MOTA). This metrics takes several factors into account: (i) false positives, i.e., erroneously detected objects, (ii) false negatives, i.e., objects missed by the tracker, and (iii) mismatches, i.e., objects that were assigned a new identificatory when switching frames. The MOTA score can be calculated using equation (3), where *FPt* is the number of false positives in frame *t*, *FNt* the number of false negatives in frame *t*, *mmet* the number of mismatches in frame *t*, and *gt* the number of objects in frame *t*.

(3)

To compare the performance of a technology against the reference the well-known Bjøntegaard delta bit rate (BD-rate) metric is used, which was first introduced in [1]. Information on how BD-rate is calculated can be found in [2]. One major difference to [2] is that instead of PSNR as the distortion metric the appropriate machine task metric is used, i.e., for object detection and instance segmentation mAP, and for object tracking MOTA.

More details about the evaluation process for the study of optimizations for encoders and receiving systems for machine analysis used by JVET can be found in the related document on common test conditions [4].

# Pre-processing technologies

There are various technologies that can be applied to the input video. While some technologies alter the source, others do not change the pixel values and only extract information that the encoder later can use to improve the encoding performance.

## Region of Interest-based methods

One often-used optimization method is region of interest (RoI)-based coding. Here the input video is analyzed in some way and the result of this analysis is made available to the encoder, which in turn can optimize the encoding towards machine consumption. The analysis can be done using various methods, e.g., neural networks. An example of a pipeline that can be used for RoI-based approaches is shown in Figure 3.

Input video

Analysis

Encoder

Decoder

Machine consumption

Figure 3. Pipeline for RoI-based systems

In one implementation example, an object detection network is used to analyze the input data. This network produces a list of objects that can be found in the current frame. This list can then be used by the encoder for example to separate foreground and background with the purpose of encoding the foreground at a better quality and the background at a lower quality. One such encoding method is described in clause 8.1. In this example, the analysis does not change the input video, which is forwarded without any changes to the encoder.

In other RoI-based methods the pre-processing might change the input video, for example by applying different pre-processing methods on the foreground and background, or specific parts of the video, such as the background, may be subsampled. Exemplar methods of processing foreground and background differently are given below.

## Pre-processing methods of foreground and background

After pre-analysis that determines the foreground and background areas, one straightforward way to handle the background that is less critical for machine consumption is to “eliminate” it by setting the corresponding sample values in the background to a constant value. However, some portions of background, for example those immediately surrounding the foreground area, could still be useful for machine vision tasks. Therefore, machine vision-related background regions can be preserved to a certain extent with low-pass filtering, such as a Gaussian filter with a sliding window, where the window size can be set based on the input video resolution.

During pre-analysis stage, extracted features can reveal refined importance information of the input video. In other words, compared with binary classification of foreground vs. background, these extracted features can provide importance information at a finer granularity. Therefore, such extracted features can be used to determine how to process foreground and background differently.

## Temporal subsampling

In some use cases, for example when the frame rate is high, a way to reduce the bit rate without a strong negative impact on the machine consumption performance is to skip certain frames and encode the video at a lower frame rate. One example is to remove every other frame from the input video and encode the video at half frame rate. This can be done in a dynamic manner, for example by evaluating the motion between two or more frames and if there is only little motion, a frame can be removed. In some cases, if the receiving system requires a specific frame rate, a corresponding post-processing technology that upsamples to full frame rate needs to be applied.

## Spatial subsampling

If the analysis of a video shows that it contains primarily large objects, one way to increase the BD-rate performance can be to perform spatial subsampling on the input video. This will result in fewer pixels in subsampled frames to encode, and thus likely faster encoding and bit rate savings. The spatial subsampling can be applied adaptively, for example, based on the diversity characteristics of the visual content and the target bit rate. For example, depending on if the input video is captured by regular camera as natural scenes, or if the input video is captured by infrared sensor as thermal images, different spatial subsampling methods can be applied.

# Encoding technologies

Some technologies can directly influence how the encoder performs and change the distribution of bits on either a frame-level or a sequence-level.

## Quantization Parameter adaption for Region of Interest coding

One method that is available in many video compression standards is adaptive quantization parameter (QP). Here the encoder can change the QP on a sub-frame level to optimize the encoding for the application. Due to this versatility, adaptive QP can be used in many different use cases to improve performance without any changes to the decoding process.

The decision on where to change the QP by how much can be made by the encoder based on an analysis of the input video. Another option is to utilize the output of an external analyzer such as described in clause 7.1. In this case, the encoder receives information about the positions of objects in each frame to make a differentiation between foreground and background.

One option is to use the base QP of the frame for areas that contain objects, i.e., foreground areas, and increase the QP for background areas, resulting in fewer bits being used to encode the background. As the background is not needed for a good machine consumption performance, this is a straight-forward way to reduce the bit rate. As an extension, it can also be beneficial to encode large objects with slightly higher QPs. As it is generally easier to detect larger objects, reducing the bit rate for large objects usually does not reduce the performance of the machine consumption task.

Utilizing the analysis, complexity can be traded against bitrate. If a light-weight neural network is used to perform the analysis, the encoder is recommended to take into account that not all relevant objects may be found and thus it could be detrimental to reduce the quality for the background too much as there might be objects that the initial analysis missed. These objects might still be important for the machine consumption task and if the background is encoded in sufficient quality, the machine consumption network has a chance of detecting objects in the background even if they are coded in lower quality.

If the encoding system has a lot of resources, it can employ a neural network of higher complexity for the analysis. With a better and more certain analysis, the bit rate for the background can be reduced more as there are likely fewer objects that have been missed in the initial analysis.

## Quantization step changes for temporal layers

Commonly used encoder configurations place different frames on different temporal layers, i.e., assigning different temporal identifiers (TID). This has the purpose to create hierarchical structures that indicate from which previously coded frames the encoder can create predictions for the current frame. One aspect of these hierarchical structures is that frames cannot be referenced by other frames on a higher temporal layer. This way not every decoded picture needs to be stored in the decoded picture buffer. Another aspect of the hierarchical structure is that frames on higher temporal layers can be encoded with higher QPs, i.e., lower quality, as they will be not or more rarely used as references by other frames. An example of a hierarchical structure is shown in Figure 4.

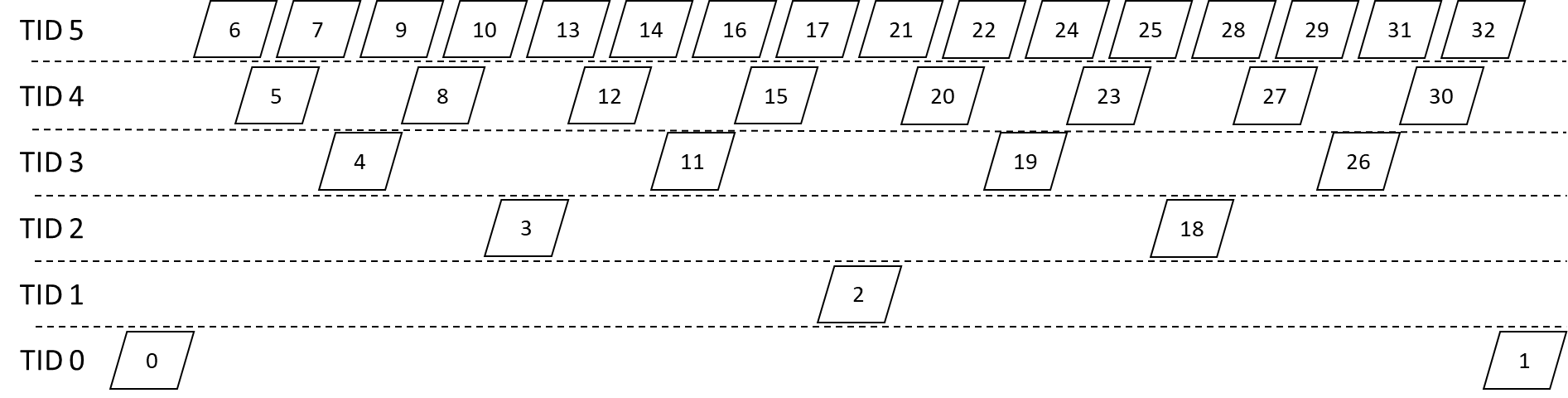


Figure 4. An example hierarchical referencing structure in a VVC random access bitstream with a total length of 33 frames. The display order is left-to-right, and the numbers specifies the order of the coded pictures in the bitstream.

This hierarchy of frames can be exploited by encoding frames on higher temporal layers using higher QPs, i.e., reducing the number of bits spent on these frames in total. This is also done in the case of the common test conditions for standard definition range for VVC. In the use case of coding video content for machine consumption this characteristic can further be exploited by increasing the QP for frames on higher temporal layers more. Due to the motion compensation included in video coding standards not many bits are needed to compress frames on high temporal layers while still being able to reconstruct high quality B‑frames.

# Post-processing technologies

There are various technologies that can be applied on the decoded video to make the post-processed video more suitable and better performing for machine consumption.

## Temporal resampling

Temporal resampling or frame reconstruction can for example be performed using interpolation filters or neural networks specifically trained for this purpose.

## Spatial resampling

Spatial resampling can for example be performed using conventional filters (e.g., VVC interpolation filter, VVC reference picture resampling filter) or neural networks specifically trained for this purpose.

# Bibliography

1. G. Bjøntegaard, “Calculation of average PSNR differences between RD-curves,” in ITU-T SG 16 Q.6 document VCEG-M33, 13th VCEG meeting, Austin, Texas, USA, Apr. 2001 (available online at <https://www.itu.int/wftp3/av-arch/video-site/>).
2. ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, “Working practices using objective metrics for evaluation of video coding efficiency experiments,” ITU-T HSTP-VID-WPOM | ISO/IEC TR 23002-8:2021.
3. C. Hollmann, J. Ström, P. Wennersten, L. Litwic, R. Sjöberg, M. Damghanian, ”Information about Ericsson’s response to the CfP for Video Coding for Machines,” Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 input document JVET-AB0275, 28th JVET meeting, by teleconference, October 2022 (available online at <https://jvet-experts.org/>).
4. ISO/IEC JTC 1/SC 29/WG 5, “Common test conditions for optimization of encoders and receiving systems for machine analysis of coded video content,” Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 output document JVET-AC2031, 29th JVET meeting, by teleconference, January 2023 (available online at <https://jvet-experts.org/>).
5. ISO/IEC JTC 1/SC 29/WG 2, “Call for Evidence for Video Coding for Machines,” MPEG Technical Requirements output document w20126, 133rd MPEG meeting, January 2021 (available online at <https://www.mpeg.org/structure/technical-requirements/>)
6. ISO/IEC JTC 1/SC 29/WG 2, “Call for Proposals for Video Coding for Machines,” MPEG Technical Requirements output document w21546, 138th MPEG meeting, April 2022 (available online at <https://www.mpeg.org/structure/technical-requirements/>)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_