**** **ISO/IEC JTC 1/SC 29/WG 2 N41**

**ISO/IEC JTC 1/SC 29/WG 2**

**MPEG Technical requirements   
Convenorship: SFS (Finland)**

**Document type:** Output Document

**Title:** **Evaluation Framework for Video Coding for Machines**

**Status:** Approved

**Date of document:** 2021-01-15

**Source:** ISO/IEC JTC 1/SC 29/WG 2

**Expected action:** none

**Action due date:** none

**No. of pages:** 48 (Including the cover page)

**Email of Convenor:** igor.curcio@nokia.com

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg2>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 2**

**MPEG TECHNICAL REQUIREMENTS**

**ISO/IEC JTC 1/SC 29/WG 2 N41**

**Online – January 2021**

**Title: Evaluation Framework for Video Coding for Machines**

**Source: WG 2 MPEG Technical requirements**

**Status: Approved**

**Serial number: 20125**

**Editor(s): M. Rafie (GTI),**

**Yuan Zhang (China Telecom),**

**Shan Liu (Tencent)**

**Table of Contents**

[1 Introduction 4](#_Toc61526194)

[2 Test Conditions 4](#_Toc61526195)

[3 Anchor Definition and Requirements 4](#_Toc61526196)

[3.1 NN Tasks …………………. 5](#_Toc61526197)

[3.2 Network architectures 6](#_Toc61526198)

[3.3 Datasets …………………. 6](#_Toc61526199)

[3.3.1 Recommended Datasets for MPEG-VCM Anchor Generation 7](#_Toc61526200)

[3.3.2 Datasets for Testing and Evaluation of VCM Technology 8](#_Toc61526201)

[3.3.3 Additional Datasets for Testing and Evaluation of VCM technology 11](#_Toc61526202)

[4 Evaluation Methods and Procedures 11](#_Toc61526203)

[4.1 Proposed Processing Pipelines 12](#_Toc61526204)

[4.2 Pre/Post-Processing Data (image) Conversion 12](#_Toc61526205)

[4.2.1 Processing Pipeline for Downscaled Resolution 13](#_Toc61526206)

[4.3 Anchor Reference Compression / Decompression 14](#_Toc61526207)

[4.4 Anchor Performance Curve Generation 14](#_Toc61526208)

[4.4.1 Anchor Curve Generation 14](#_Toc61526209)

[4.4.2 Anchor Metrics 15](#_Toc61526210)

[5 Evaluation Approach for Machine and Human Consumption 15](#_Toc61526211)

[5.1 Complexity Measurement 17](#_Toc61526212)

[6 References 18](#_Toc61526213)

[7 Appendix A: Anchor Generation for CfE 19](#_Toc61526214)

[7.1 Object Detection ........... 19](#_Toc61526215)

[7.1.1 Nokia vs. Tencent (Open Images) 19](#_Toc61526216)

[7.1.1 Konkuk Univ vs. ETRI (FLIR) 21](#_Toc61526217)

[7.1 Object Segmentation (Instance) 25](#_Toc61526218)

[7.1.1 Nokia vs. ETRI (Open Images) 25](#_Toc61526219)

[7.2 Object Tracking (HiEve-10) 27](#_Toc61526220)

[7.2.1 SJTU vs China Telecom 27](#_Toc61526221)

[7.3 Action Recognition (HiEve-10) 33](#_Toc61526222)

[7.3.1 SJTU vs China Telecom 33](#_Toc61526223)

[7.4 Pose Estimation (HiEve-10) 39](#_Toc61526224)

[7.4.1 SJTU vs China Telecom 39](#_Toc61526225)

[8 Appendix B: Anchor Runtime Measurement 45](#_Toc61526226)

[9 Appendix C: Anchor Metrics 46](#_Toc61526227)

[9.1 Task: Object Detection / Object Segmentation 46](#_Toc61526228)

[9.2 Task: Object Tracking 46](#_Toc61526229)

[9.3 Definition of Metrics …. 47](#_Toc61526230)

# Introduction

The MPEG activity on Video Coding for Machines (VCM) aims to standardize a bitstream format generated by compressing either a video stream or previously extracted features. The bitstream should enable multiple machine vision tasks. VCM shall be able to

* Efficiently compress the bitstream; the size of the compressed features shall be less than the encoded video stream using state-of-the-art video compression technologies such as VVC.
* Use a common bitstream to support single or multiple tasks. The decompressed bitstream should be general enough to be used for different scenarios, for example, object detection and segmentation.
* Support varying performance for multiple tasks as measured by the appropriate metrics. This performance level may depend on the application.
* For hybrid machine/human vision use cases, a common bitstream should be used for machine and human consumption; additional bitstream(s) is optional for the reconstruction of the compressed bitstream for human consumption
* The bitrate of the additional compressed bitstream shall be less than the bitrate of the bitstream at similar quality as measured by PSNR, which is the output of the VVC encoding of the unprocessed video.

MPEG VCM has identified a set of relevant use cases and related requirements [1], focusing on the machine-vision; and the hybrid machine/human-vision use cases.

* Datasets: which datasets should be used for which sub-tasks, where these datasets can be obtained, how the datasets are split into training and validation data
* Metrics: which metric shall be used for which sub-tasks, how these metrics are calculated, what to compare performance results against

# Test Conditions

Decoded video/feature shall be tested for one or more key tasks for a specific use case and compare the performance results to current anchors. Retraining the shared backbone is permitted using joint training or other approaches in the case of two or more key tasks. Modifications and training of the task-specific networks are allowed but need to be reported in detail. In some cases, the encoder may know the task-specific neural networks at the decoder side. In this document, framework refers to the used datasets and software packages.

# Anchor Definition and Requirements

VVC/H.266 codec with software version VTM-8.2 is used as the reference for the performance evaluation of the MPEG-VCM encoder. Table 1 shows the tasks considered in the Call for Evidence (CfE) along with their metrics, datasets, benchmarks, and training/testing description.

Table 1 *Training and test conditions, key metrics, datasets, benchmarks for various tasks*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | **Metrics** | **Datasets** | **Benchmarks** | **Training/Testing** |
| Object Detection | [mAP](https://mc.ai/the-confusing-metrics-of-ap-and-map-for-object-detection/)  vs  BPP/Rate  PSNR | Open Images  Compressed  (Image) | <https://storage.googleapis.com/openimages/web/index.html> | Version 6. |
| FLIR Thermal Dataset  [Compressed]  (Image) | <https://www.flir.com/oem/adas/adas-dataset-form/> | Ability to sense thermal infrared radiation or heat |
| Object (Instance) Segmentation | [mAP](https://mc.ai/the-confusing-metrics-of-ap-and-map-for-object-detection/)  vs  BPP/Rate | Open Images  Compressed  (Image) | https://storage.googleapis.com/openimages/web/index.html | Version 6. |
| Object Tracking | MOTA/PSNR  vs  Rate/BPP | HiEve-10  [Uncompressed]  (video) | <http://humaninevents.org/> | SJTU - Human in Event  HiEve-10 has 7 training videos and 3 testing videos |
| Action Recognition | Frame-mAP/PSNR  Vs  Rate/BPP | HiEve-10  [Uncompressed]  (video) | <http://humaninevents.org/> | SJTU - Human in Event  HiEve-10 has 7 training videos and 3 testing videos |
| Pose Estimation | AP/PSNR  vs  Rate/BPP | HiEve-10  [Uncompressed]  (video) | <http://humaninevents.org/> | SJTU - Human in Event  HiEve-10 has 7 training videos and 3 testing videos |
| Hybrid Machine / Human Vision | BD-PerVal  (mAP/MOTA, SSIM, PSNR vs. BPP/Rate) | Any combination of listed Datasets  (video) | See above | See above |

## NN Tasks ………………….

For the MPEG-VCM performance evaluation, the anchors are generated for the following key NN tasks (see Appendix A):

* Object Detection (still image)
  + Open Images
    - Nokia vs. Tencent (crosschecker)
  + FLIR
    - Konkuk Univ vs. ETRI (crosschecker)
      * Fine-tuned (mAP/bpp) (Thermal Images to be used for CfE)
      * Machine/human vision task (SSIM and PSNR)
* Object Segmentation (Instance) (still image)
  + Open Images
    - Nokia vs. ETRI (crosschecker)
* Object Tracking
  + HiEve-10
  + SJTU vs. China Telecom (crosschecker)
* Pose Estimation
  + HiEve-10
  + SJTU vs. China Telecom (crosschecker)
* Action Recognition
  + HiEve-10
  + SJTU vs. China Telecom (crosschecker)

## Network architectures

* Object Detection:
  + Faster R-CNN X101-FPN (part of Facebook AI Research’s Detectron)

* Object Segmentation (Instance):
  + R50-FPN (part of Facebook AI Research’s Detectron)

* Object tracking:
  + JDE-1088x608
* Action Recognition:
  + Slowfast
* Pose Estimation:
  + HRNet
* Optional requirement
  + As additional performance data, other network architectures are allowed to be used per task. However, it is up to the proponents to provide comparable performance data with the anchors specified in this document

## Datasets ………………….

The datasets to be used for anchor generation evaluation framework and VCM standardization process shall be of high quality, available and downloadable, sufficiently modern, have permissive licenses, and support adequate pre-trained models.

* Datasets must be of high quality such that a VVC encoder can create several coded versions of the sequence with noticeable degradation in quality. It is recommended to use uncompressed datasets where possible.
* In some datasets, most images are compressed with quality factors around 96 and 90, while some images were even compressed with quality factor around 80. It is recommended to build up a subset of the dataset by removing the lower quality images.
* Datasets shall be capable of generating anchors per requirements
* Datasets shall be available and downloadable
* Datasets shall be sufficiently modern
* Datasets should support pre-trained models

Proponents are invited to look into the datasets and raise a flag in case of encountering issues such as copyrights, etc. (Note: It is intended to have these datasets to be downloadable from the MPEG site). Table 2 shows the recommended datasets for VCM anchor generation.

### Recommended Datasets for MPEG-VCM Anchor Generation

The following are identified as the datasets with permissible licenses are recommended for the MPEG-VCM anchor generation of CfE (see Table 2):

* Open Images v6
* FLIR
* HiEve-10
* SFU-HW-Objects-v1
* Tencent Video Dataset (TVD)

Currently, the relevant files/data corresponding to Open Images and FLIR are uploaded at “MPEG Content Repository” (<https://mpegfs.int-evry.fr/>):

* username = sc29wg11
* Then, log into the mpeg content repository using username = mpegcontent
* Subdirectories: Explorations/VCM
* VCM path: <https://mpegfs.int-evry.fr/mpegcontent/ws-mpegcontent/Explorations/VCM/>

Table 2*. The following datasets are recommended to be used for MPEG-VCM anchor generation of CfE and CfP.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Image /**  **Video** | **Compressed /**  **Uncompressed** | **Link Location** | **Comments** |
| Open Images v6 | Image | Compressed | <https://storage.googleapis.com/openimages/web/index.html> | * 9M images, 600 classes, Modern. * 16M bounding boxes for 600 object classes on 1.9M images * Listed as using CC BY 2.0 and CC BY 4.0 Licensing terms |
| TVD | Image / video | Uncompressed | MPEG-VCM Depository | * Test sequences for evaluation of VCM machine tasks, i.e., object detection, instance segmentation or object tracking, etc. * Each clip contains 128 frames * Resolution of the raw sequences is 3840x2160 * Tencent will label these sequences, provide anchors results * Granted permissible license for VCM group. |
| FLIR | Image | Compressed | <https://creativecommons.org/licenses/by/4.0/>  <https://www.flir.com/oem/adas/adas-dataset-form/> | * Ability to sense thermal infrared radiation or heat * 1.1 License * Limited, revocable, non-exclusive, non-sublicensable, non-transferable license to access and use the Image Data in the field of neural network development for automotive and other autonomous vehicle applications, and for general non-commercial educational and research purposes. * Shall be used for the Image Data / ADAS exclusively for the scientific research, testing and evaluation purposes of the MPEG VCM activity and for the presentation of results inside MPEG (“Additional Approved Use”). * Provided under the Creative Commons license BY 4.0 (CC BY 4.0) |
| HiEve-10 | Video | 10 Uncompressed | <http://humaninevents.org/> | * SJTU - Human in Event * HiEve-10 has 7 training videos and 3 testing videos * Except for these 10 videos, other videos in this Content can only be used for non-commercial purposes. * Proponent agreed to provide permission (permissible license) for MPEG-VCM activities at MPEG131 (2020-01-08 Received) |
| SFU-HW-Objects-v1 | Video | Uncompressed | <https://dx.doi.org/10.25314/7d8efc0a-3943-4738-b7a5-72badb04d765> | * Labeled video data * Already being used for MPEG (HEVC) – appears to be fine for standardization activity * The dataset is provided under the Creative Commons license BY 4.0 (CC BY 4.0) * Proponent agreed to provide permission (licensing free for labels) for MPEG-VCM activities at MPEG131 (by 2020-07-03) |

### Datasets for Testing and Evaluation of VCM Technology

In addition to the MPEG-VCM permissible datasets specified in Table 2 for the anchor generation of CfE, Table 3 provides additional datasets relevant to the tasks considered in this document for further testing and evaluation of VCM technology.

Table 3*. Additional datasets relevant to the tasks considered in this document for further testing and evaluation of VCM technology*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Image /**  **Video** | **Compressed /**  **Uncompressed** | **Link Benchmarks** | **Comments** |
| COCO | Image | Compressed | <http://cocodataset.org/#detection-leaderboard> | For COCO, use 2017 Val set for evaluation and 2017 Train in the case of retraining  80 object classes  “The annotations in this dataset along with this website belong to the COCO Consortium and are licensed under a [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/legalcode).” (see Note 1.)  “The COCO Consortium does not own the copyright of the images. The use of the images must abide by the [Flickr Terms of Use](https://www.flickr.com/creativecommons/). The users of the images accept full responsibility for the use of the dataset, including but not limited to the use of any copies of copyrighted images that they may create from the dataset.” |
| CityScapes | Image | Uncompressed | <https://www.cityscapes-dataset.com/benchmarks/> | For Cityscapes, use defined train and validation sets  “This Cityscapes Dataset is made freely available to academic and non-academic entities for non-commercial purposes such as academic research, teaching, scientific publications, or personal experimentation.” |
| CityPersons | Image | Uncompressed | <https://www.cityscapes-dataset.com/benchmarks/> | For Citypersons, use defined train and validation sets  “CityPersons, a new set of person annotations on top of the Cityscapes dataset.” for pedestrian detection” |
| [KITTI](http://www.cvlibs.net/datasets/kitti/index.php) | Image | Uncompressed | <http://www.cvlibs.net/datasets/kitti/eval_object.php> | We recommend using the predefined splits  “All datasets and benchmarks on this page are copyright by us and published under the [Creative Commons Attribution-NonCommercial-ShareAlike 3.0](http://creativecommons.org/licenses/by-nc-sa/3.0/) License. This means that you must attribute the work in the manner specified by the authors, you may not use this work for commercial purposes and if you alter, transform, or build upon this work, you may distribute the resulting work only under the same license.” |
| DAVIS 2016 / 2017 | Video | Compressed | <https://davischallenge.org/> | We recommend using the semi-supervised mode for higher accuracy.  Creative Commons Attributions 4.0 License (non-commercial use) |
| MOT20 | Video | Compressed | <https://arxiv.org/pdf/1906.04567.pdf> | Dataset split is available from the Tracking Challenge, available on their website. |
| UCF101 | Video | Compressed | <https://www.crcv.ucf.edu/data/UCF101.php> | Action recognition  13320 videos from 101 action categories. |
| HiEve | Video | 14 Uncompressed | <http://humaninevents.org/data.html?title=1> | Human in Event  19 training videos, 13 test videos, and a total of 32 videos  No issues with licensing – can provide license/permission to MPEG-VCM  Annotation of the bounding box, ID, key points and behavior, dense pose. Can be used for multi-target tracking  Proponent agreed to provide permission (permissible license) for MPEG-VCM activities at MPEG131 (by 2020-07-03 - Received) |
| SFU-HW-Objects-v1 | Video | Uncompressed |  | Labeled video data  Object labeled dataset on raw video sequences  Already being used for MPEG (HEVC) – appears to be fine for standardization activity  Can be used for compression and object detection simultaneously  Need to be investigated to see whether it large enough to be used for VCM activities  The dataset is provided under the Creative Commons license BY 4.0 (CC BY 4.0)  Proponent agreed to provide permission (licensing free for labels) for MPEG-VCM activities at MPEG131 (by 2020-07-03) |
| Open Images | Image | Compressed | <https://storage.googleapis.com/openimages/web/index.html> | 9M images, 600 classes, Modern.  It contains a total of 16M bounding boxes for 600 object classes on 1.9M images, making it the largest existing dataset with object location annotations.  Listed as using CC BY 2.0 and CC BY 4.0 Licensing terms |
| FLIR Thermal Dataset | Image | Compressed | <https://www.flir.com/oem/adas/adas-dataset-form/> | Ability to sense thermal infrared radiation or heat  “1.1 License. As between you and us, we or our licensors own and reserve all right, title, and interest in and to the Image Data. We grant you a limited, revocable, non-exclusive, non-sublicensable, non-transferable license to access and use the Image Data in the field of neural network development for automotive and other autonomous vehicle applications, and for general non-commercial educational and research purposes. Except as provided in this Section 1, you obtain no rights under this Agreement from us or our licensors to the Image Data, including the right to reproduce, redistribute or make derivative works of the Image Data” |
| TVD | Image / video | Uncompressed | MPEG-VCM Depository | * Tencent / Tsinghua University * Test sequences for evaluation of VCM machine tasks, i.e., object detection, instance segmentation or object tracking, etc. * Each clip contains 128 frames * Resolution of the raw sequences is 3840x2160 * Tencent will label these sequences, provide anchors results * Will grant permissible license for VCM group. |

### Additional Datasets for Testing and Evaluation of VCM technology

In consideration of the datasets adopted for MPEG-VCM, additional datasets were identified from the website [www.datasetlist.com](http://www.datasetlist.com) for consideration by the VCM group.

The aggregated datasets were filtered to commercial licenses only, and to the categories Image, Self-driving, and Medical. The filtered datasets were investigated further.

It is recommended that to ease uncertainty on accessibility, datasets with permissive licenses for commercial use are adopted where possible for testing and evaluation in CfE, CfP, and standardization processes.

# Evaluation Methods and Procedures

The evaluation procedure and metrics are described in section 2 above. The metrics consist of two parts, one relating to feature extraction and one relating to compression of processed or unprocessed video. The metrics and anchors for feature extraction will be considered later.

The majority of these datasets have publicly defined training and validation sets. In the case this is not available, we will release a training and testing split for comparison. This list is not exhaustive, and proponents are free to use their own datasets for each of the key tasks.

The input images and labels for training and testing are directly taken from the dataset for specific use cases as listed above. This leads into a general feature extractor such as a convolutional neural network, which converts the images or video into a stream of processed video. The resulting features are then fed into different machines, whose results are calculated with respect to the appropriate metric. Proponents are asked to report this result along with the current state of the art on the chosen group of tasks, which will be released by MPEG-VCM. A comparison will be made regarding the performance across the different tasks in the group measured by the relevant metric.

Regarding the compression of processed or unprocessed video, proponents are asked to test the compression ratio on the processed or unprocessed video. This compression ratio should be given as a comparison to the released compression ratio of VVC on the unprocessed video.

For human consumption use cases, proponents shall report BD-rate curves. BD-rate should be calculated in the way as other standardization groups, e.g. JVET. The performance reporting format is specified in the MPEG-VCM BD-rate and BD-AP/BD-Accuracy reporting template [2]

* Use case-specific performance metrics, with the key tasks and metrics as defined in Table 1. Proponents shall perform the evaluation themselves, with the experimental conditions described in [1].
* Compression efficiency, runtime complexity, and memory consumption of compression/ decompression (measurement is independent of the use case). Proponents shall perform the evaluation themselves based upon a provided unprocessed or processed video. In the case of processed video, the output may come from a common neural network or general feature extraction methods regarding the specific key tasks. As an example, these common neural network backbones may be VGG, ResNet, Inception and the specific frameworks depend on the key tasks. For detection and segmentation, an example may be Mask R-CNN or YOLO.

## Proposed Processing Pipelines

Possible pipeline architectures for technology proposal are shown in Figures 1.a, 1.b, and 1.c. Anchors are generated on the pipeline architecture shown in Figure 1.a.

|  |
| --- |
|  |
| 1. Pipeline 1 |
|  |
| 1. Pipeline 2 |
|  |
| 1. Pipeline 3 |

Figure **1**. **Proposed processing pipeline**

## Pre/Post-Processing Data (image) Conversion

For input data processing, it is suggested to use FFmpeg release 4.2.2. FFmpeg can be used for data format conversion, up/down-sampling, and resizing (cropping/padding/scaling) the image.

* + FFmpeg 4.2.2
    - Resolution: Scaling / resolution (100%, 75%, 50%, 25%):
* Scale PNG image to new resolution:

ffmpeg -i input.png -vf “scale=NEW\_WDT:NEW\_HGT“ output.png

for 100%: -vf “pad=ceil(iw/2)\*2:ceil(ih/2)\*2”

for 75%: -vf "scale=ceil(iw\*3/8)\*2:ceil(ih\*3/8)\*2"

for 50%:   -vf "scale=ceil(iw\*/4)\*2:ceil(ih\*/4)\*2"

for 25%: -vf "scale=ceil(iw\*/8)\*2:ceil(ih\*/8)\*2"

* + - Format conversion: Convert PNG 🡨🡪 YUV:

ffmpeg -i input.png -f rawvideo -pix\_fmt yuv420p *-dst\_range 1* output.yuv

ffmpeg -f rawvideo -pix\_fmt yuv420p10le -s WDTxHGT -*src\_range 1 -i* input.yuv -frames 1 -pix\_fmt rgb24 output.png

* + - Scale image to original image size:

For 100%: ffmpeg -i input.png -vf “crop=ORIG\_WDT:ORIG\_HGT“ output.png

Other resolutions: ffmpeg -i input.png -vf “scale=ORIG\_WDT:ORIG\_HGT“ output.png

### Processing Pipeline for Downscaled Resolution

There are two possible processing pipelines for the downscaled ground-truth image as are shown in Figure 1. For anchor generations, an alternative processing pipeline shown in Figure 2.b is recommended.

|  |
| --- |
|  |
| 1. Original Pipeline |
|  |
| 1. Alternative pipeline – Upscaled decoded image |

Figure 2. Possible pipelines for downscaled resolutions

## Anchor Reference Compression / Decompression

VTM8.2 encoder is used to compress generated YUV files under the default “All Intra” configuration and ConformaceWindowMode set to 1. ConformaceWindowMode is equal to 1 represent padding width and height of processing image automatic to multiple of minimal CU size.

A total of six evenly-spaced QPs are recommended, respectively (17 optional), 22, 27, 32, 37, 42, and 47.

* VVC: Reference software VTM-8.2
* JVET Common Test Conditions (CTC-420) with Random Access (RA) condition for videos
* JVET Common Test Conditions (CTC-420) with All Intra AI condition for images
* encoder -c cfg/encoder\_intra\_vtm.cfg -i input.yuv -o reconstruct.yuv -b compress.vvc -q QP --ConformanceWindowMode=1 -wdt WDT -hgt HGT -f 1 -fr 1 --InternalBitDepth=10
* decoder -b compress.vvc -o decode.yuv

## Anchor Performance Curve Generation

### Anchor Curve Generation

To generate the performance curves of the anchors, mAP/MOTA vs BPP/bitrate curves are to be produced with the following specifications.

* + Specify max. endpoint as the uncompressed performance (mAP/MOTA)
  + Specify min. endpoint performance (mAP/MOTA)
  + Generate performance (mAP/MOTA v.s. BPP) curves for each task based on the 4x resolutions and QPs within the range of min/max endpoints
  + Specify the minimum threshold (i.e., below which the performance results are not acceptable). There are five options for min. threshold [2]:
    - -10%: The lower threshold is set to 10 percentage points below the uncompressed performance.
    - -20%: The lower threshold is set to 20 percentage points below the uncompressed performance.
    - Half: The lower threshold is set to half of the uncompressed performance.
    - None: No lower threshold is used.
    - Fixed threshold – users can specify an arbitrary (fixed) min. threshold.
  + The performance curve is the Pareto-Front created from the generated 4 curves above.

### Anchor Metrics

A more detailed description of the accuracy measurement metrics refer to Appendix C.

* Accuracy measurement – mAP/MOTA is used to measure the accuracy performance
  + mAP, mAP@0.5, [AP@[0.5:0.95](mailto:AP@[0.5:0.95)]
  + MOTA
* Compression efficiency measurement – BPP/bitrate is used to measure the cost for storage/transmission of the generated bitstream for VCM
  + BPP calculation: BPP is calculated with respect to the original image resolution (not the downscales image)
* For machine/human-vision tasks, BD-rate curves (mAP/MOTA, SSIM, PSNR vs. BPP/bitrate) should be calculated in the way as other standardization groups, e.g. JVET.
* The performance curves shall be compared against the anchor performance curves using the BD-rate, BD-mAP, BD-MOTA, and BD-PSNR functions in the MPEG-VCM reporting template [2].

# Evaluation Approach for Machine and Human Consumption

The evaluation process of the video test data set for machine consumption and human consumption using VVC as an anchor is shown in Figure 3 and Figure 4, respectively.

|  |
| --- |
|  |

Figure 3. Evaluation approach of video test data using VVC anchor - machine consumption

|  |
| --- |
|  |

Figure 4. Evaluation approach of video test data using VVC anchor - human consumption

The feature data type and format information are beneficial for compression experts to know the properties and limitations of the feature data types and formats to increase the quality of their proposals. The list of feature data types and formats is recommended to be as exhaustive as possible and to include all relevant information such as allowed values and data ranges.

Table 4 shows an overview of different data types for different tasks required by the use cases described in [1].

Table 4. *Feature data types and their description for various tasks*

|  |  |  |
| --- | --- | --- |
| **Task** | **Type of data** | **Description** |
| Object detection | List of bounding boxes | Maximum number of bounding boxes: TBD  Each bounding box has four attributes:   * pos\_x (integer): offset from left picture edge: 0 to MAX\_PIC\_WIDTH * pos\_y (integer): offset from top picture edge: 0 to MAX\_PIC\_HEIGHT * size\_x (integer): width from left edge of bounding box: 1 to MAX\_PIC\_WIDTH * size\_y (integer): height from top edge of bounding box: 1 to MAX\_PIC\_HEIGHT |
| Object segmentation | Matrix | Matrix size: INPUT\_WIDTH x INPUT\_HEIGHT  All elements of the matrix are either a single integer value or a list of three integer values (for different color formats). The range of the values depends on the chosen bit depth. |
| Object tracking | List of bounding boxes | Maximum number of bounding boxes: TBD  Each bounding box has five attributes:   * pos\_x (integer): offset from left picture edge: 0 to MAX\_PIC\_WIDTH * pos\_y (integer): offset from top picture edge: 0 to MAX\_PIC\_HEIGHT * size\_x (integer): width from left edge of bounding box: 1 to MAX\_PIC\_WIDTH * size\_y (integer): height from top edge of bounding box: 1 to MAX\_PIC\_HEIGHT   box\_id (integer): identifier for each box to allow tracking through multiple frames |

## Complexity Measurement

Encoder/decoder runtime has served as a reasonable proxy for computational complexity in JVET / VVC standardization projects. It is recommended the runtime metric to be reported for anchors generation in Machine vision and Human vision tasks.

The complexity of the codec shall be independent of pre-processing methods and NN tasks used. For the recommended method to measure the codec complexity see Appendix B.

# References

|  |  |
| --- | --- |
| [1] | MDS19841\_WG02\_N00018, "Use cases and draft requirements for Video Coding for Machines," Online, Oct 2020. |
| [2] | C. Hollmann, (Ericsson), S. Liu, W. Gao, X. Xu and (Tencent), "m56252\_[VCM] On VCM Reporting Template," Jan 2021. |
| [3] | M. Everingham, V. Gool, I. Williams C and et al., "The Pascal Visual Object Classes (VOC) Challenge[J]," *International Journal of Computer Vision,* pp. 88(2):303-338., 2010. |
| [4] | M. Everingham, A. Eslami S M, L. Van Gool and e. al., "The Pascal Visual Object Classes Challenge: A Retrospective[J].," *International Journal of Computer Vision,* pp. 111(1):98-136., 2015. |
| [5] | K. Bernardin, A. Elbs and A. Stiefelhagen, "Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment [J]," *Sixth IEEE International Workshop on Visual Surveillance in Conjunction with Eccv,* May, 2008. |
| [6] | A. Horé and D. and Ziou, "Image Quality Metrics: PSNR vs. SSIM," *2010 20th International Conference on Pattern Recognition, Istanbul,* pp. 2366-2369, doi: 10.1109/ICPR.2010.579, 2010. |
| [7] | Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. andSimoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing,* Vols. vol. 13, no. 4, , doi: 10.1109/TIP.2003.819861, pp. 600-612, April 2004. |

# Appendix A: Anchor Generation for CfE

The following sections provide the generated anchor results for Object Detection, Object Segmentation (Instance), Object Tracking, Action Recognition, and Pose Estimation

## Object Detection ...........

### Nokia vs. Tencent (Open Images)

|  |  |  |
| --- | --- | --- |
| **Nokia, Tencent (Crosschecker)** | | |
| Network model: Faster R-CNN X101-FPN | | Dataset: COCO train2017, Open Images (Eval) |
| Test Condition: | | * Format conversion: FFmpeg 4.2.2 * Codec: VTM 8.2 * Scaling resolution: 100% * QPs: 22, 27, 32, 37, 42, 47 * Average precision: mAP@0.5 * Compression ratio: BPP |
| Processing flow for anchor generation | | |
| Test Results: | * Average precision vs. BPP * Maximum mAP (with no additional compression) for object detection task is 0.770122743 | |

|  |
| --- |
| Figure 1 - Rate-Performance of Anchors for object detection task |
| Table 1 - BPP and AP results for different QPs and resolutions in object detection task   |  |  |  | | --- | --- | --- | |  | Resolution 100% | | | QP | BPP | mAP | | 22 | 0.83601788 | 0.76517548 | | 27 | 0.496323715 | 0.76032878 | | 32 | 0.277115735 | 0.749032079 | | 37 | 0.146083233 | 0.722172818 | | 42 | 0.073909964 | 0.672795607 | | 47 | 0.035294544 | 0.566341802 | |

### Konkuk Univ vs. ETRI (FLIR)

|  |  |  |
| --- | --- | --- |
| **Konkuk Univ, ETRI (Crosschecker)** | | |
| Network model: Faster R-CNN X101-FPN | | Dataset: FREE FLIR Thermal dataset |
| Test Condition: | | * Format conversion: FFmpeg 4.2.2 * Codec: VTM 8.2 * Scaling resolution: 100%, 75%, 50% and 25% * QPs: 22, 27, 32, 37, 42, 47 * Metric: mAP, SSIM and PSNR) * Compression ratio: BPP |
| Figure 1 Processing Flow of Anchor generation for object detection on FLIR dataset | | |
| Test Results: | * Average precision, SSIM, PSNR vs. BPP * 600 images (300 RGB images and 300 IR images) from FLIR Dataset | |

|  |
| --- |
| Figure 1 - Anchor generation results of object detection using the fine-tuned network for thermal images |
| Table 1 - Anchor generation results of object detection using the fine-tuned network for thermal images   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **Scale** | **QP** | **mAP** | **Y-PSNR** | **U-PSNR** | **V-PSNR** | **SSIM** | **BPP** | | **original** |  | 40.557 |  |  |  |  | 6.861 | | **100%** | 17 | 39.279 | 47.706 | - | - | 0.997 | 2.476 | | 22 | 39.446 | 42.904 | - | - | 0.992 | 1.807 | | 27 | 39.643 | 37.954 | - | - | 0.975 | 1.224 | | 32 | 37.690 | 31.455 | - | - | 0.854 | 0.306 | | 37 | 33.784 | 29.730 | - | - | 0.796 | 0.131 | | 42 | 23.069 | 28.293 | - | - | 0.755 | 0.065 | | 47 | 10.634 | 26.580 | - | - | 0.702 | 0.030 | | **75%** | 17 | 40.229 | 26.817 | - | - | 0.894 | 1.376 | | 22 | 40.431 | 26.673 | - | - | 0.881 | 0.888 | | 27 | 39.160 | 26.262 | - | - | 0.836 | 0.400 | | 32 | 36.423 | 25.865 | - | - | 0.801 | 0.189 | | 37 | 30.992 | 25.376 | - | - | 0.768 | 0.098 | | 42 | 19.437 | 24.657 | - | - | 0.725 | 0.049 | | 47 | 7.456 | 23.673 | - | - | 0.669 | 0.022 | | **50%** | 17 | 35.955 | 26.094 | - | - | 0.838 | 0.578 | | 22 | 35.711 | 25.965 | - | - | 0.822 | 0.352 | | 27 | 34.121 | 25.726 | - | - | 0.797 | 0.193 | | 32 | 30.406 | 25.377 | - | - | 0.770 | 0.107 | | 37 | 21.838 | 24.809 | - | - | 0.734 | 0.056 | | 42 | 9.860 | 23.967 | - | - | 0.685 | 0.027 | | 47 | 2.501 | 22.935 | - | - | 0.632 | 0.012 | | **25%** | 17 | 19.172 | 24.675 | - | - | 0.743 | 0.149 | | 22 | 18.405 | 24.598 | - | - | 0.734 | 0.100 | | 27 | 16.175 | 24.433 | - | - | 0.719 | 0.064 | | 32 | 11.847 | 24.093 | - | - | 0.694 | 0.038 | | 37 | 6.142 | 23.473 | - | - | 0.657 | 0.020 | | 42 | 1.371 | 22.635 | - | - | 0.618 | 0.010 | | 47 | 0.126 | 21.705 | - | - | 0.588 | 0.005 | |

|  |
| --- |
| Figure 2 - Pareto-Front curve result of object detection using the fine-tuned network for thermal images in FLIR dataset |

## Object Segmentation (Instance)

### Nokia vs. ETRI (Open Images)

|  |  |  |
| --- | --- | --- |
| **Nokia, ETRI (Crosschecker)** | | |
| Network model: R50-FPN | | Dataset: Dataset: COCO train2017, Open Images (Eval) |
| Test Condition: | | * Format conversion: FFmpeg 4.2.2 * Codec: VTM 8.2 * Scaling resolution: 100% * QPs: 22, 27, 32, 37, 42, 47 * Average precision: mAP@0.5 * Compression ratio: BPP |
|  | | |
| Test Results: | * Average precision vs. BPP * Maximum mAP (with no additional compression) for the object segmentation task is 0.777767769 | |

|  |
| --- |
| Figure 1 - Rate-Performance of Anchors for Object Segmentation task |
| Table 1 - BPP and AP results for different QPs and resolutions in Object Segmentation task   |  |  |  | | --- | --- | --- | |  | Resolution 100% | | | QP | BPP | mAP | | 22 | 0.8276013159 | 0.772242455 | | 27 | 0.4877710640 | 0.767264709 | | 32 | 0.2720679328 | 0.755412475 | | 37 | 0.1435697441 | 0.728229707 | | 42 | 0.0725003548 | 0.675406629 | | 47 | 0.0345425040 | 0.566708183 | |

## Object Tracking (HiEve-10)

### SJTU vs China Telecom

|  |  |  |
| --- | --- | --- |
| **SJTU vs China Telecom (crosschecker)** | | |
| Network model: JDE-1088x608 | | Dataset: HiEve-10 training |
| Test Condition: | | * Format conversion: FFmpeg 4.2.2 * Codec: VTM 8.2 * Scaling resolution: 100%, 75%, 50%, 25% * QPs: 22, 27, 32, 37, 42, 47 * Metric: MOTA * Compression ratio: Bit rate |
| Test Results: | * FPS = 30 * MOTA vs bitrate * Matlab | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

## Action Recognition (HiEve-10)

### SJTU vs China Telecom

|  |  |
| --- | --- |
| SJTU vs China Telecom (Crosschecker) | |
| Network model: Slowfast | Dataset: HiEve-10, retrained model  Faster R-CNN, Bounding box detection |
| Test Condition: | * Format conversion: FFmpeg 4.2.2 * Codec: VTM 8.2 * Scaling resolution: 100%, 75%, 50% and 25% * QPs: 22, 27, 32, 37, 42, 47 * Metric: frame-mAP * Compression ratio: Bitrate |
| Test Results: | * FPS = 30 * Frame-mAP |
|  | |
|  | |
|  | |
|  | |
|  | |

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

## Pose Estimation (HiEve-10)

### SJTU vs China Telecom

|  |  |  |
| --- | --- | --- |
| **SJTU vs China Telecom (Crosschecker)** | | |
| Network model: YOLO-v3+Deep-high-resolution-network (DHRN) | | Dataset: HiEve-10  YOLO-v3 to detect bounding box |
| Test Condition: | | * Format conversion: FFmpeg 4.2.2 * Codec: VTM 8.2 * Scaling resolution: 100%, 75%, 50% and 25% * QPs: 22, 27, 32, 37, 42, 47 * Detection confidence threshold: 0.1; IOU threshold for NMS: 0.5. * Compression ratio: Bitrate |
| Test Results: | * FPS = 30 * Access to the model for training * Provided/uploaded in the MPEG-VCM repository * PyTorch | |

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

# Appendix B: Anchor Runtime Measurement

Runtime as a proxy for complexity in JVET / VVC standardization to be reported for anchors for Machine vision task and Human vision task. The proposed runtime methods for machine and human visions are:

* Machine vision tasks
  + EncTm: Time needed to convert RGB input to bitstream
  + DecTm: Time needed to convert bitstream to inference output
* Human vision
  + Encv: time needed to convert RGB input to bitstream
  + DecTv: Time needed to convert bitstream to YUV

The runtime methods for machine and human visions are shown in Figure 5 and Figure 6, respectively.

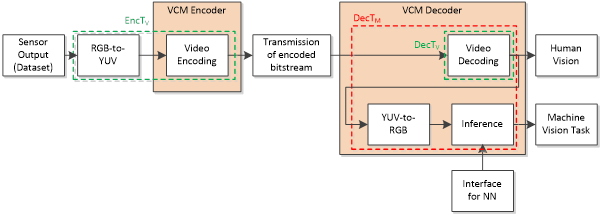


Figure 5. *Runtime measurements to be made for anchors*

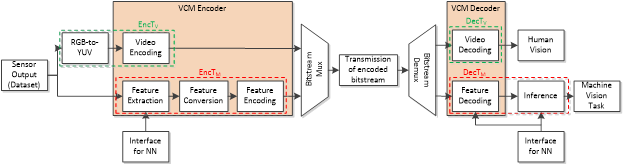


Figure 6. *Illustration of runtime measurements to be made for the proposed technology*

# Appendix C: Anchor Metrics

## Task: Object Detection / Object Segmentation

* **Metrics:** mAP (mean Average Precision) [3] [4]
* **Description:**

For a given category of object, true positive , false positive , false negative , and true negative are defined with an Intersection over Union (IoU) threshold for that category, where true/false represents the output of the neural network, positive/negative represents the label in the ground truth.

Then, recall of the given IoU threshold is defined as the proportion of all true positive examples in all true positive and false negative examples corresponding to that IoU threshold:

The precision of the given IoU threshold is the proportion of all true positive examples which are from all positive examples:

A neural network of detection or segmentation may achieve several pairs of recall and precision values corresponding to a certain IoU threshold and different confidence levels. For each recall value in the pairs, let takes the maximum precision value in all precision values for which the corresponding recall values are above the given recall value :

Average Precision (AP) of a given category of object is defined as the average value of for all recall values provided by the neural network, which can characterize the area of the entire precision-recall curve.

Mean Average Precision (mAP) is an averaged AP overall categories of objects and in a range of IoU thresholds. As an example, in MS COCO 2017 dataset, 10 IoU thresholds are taken at equal intervals from 0.50 to 0.95. In particular, AP50 and AP75 generally present the mAP when the IoU threshold is 0.50 and 0.75 respectively.

A different neural network may provide object detection or segmentation results with a different number of confidence level. This may affect the quality evaluation in VCM. So we recommend specifying the number of confidence levels provided by the neural network for each dataset/task. For example, the number of confidence levels provided by a new neural network should be the same as anchors. This will help to align potential responses to the CfE and easy comparison.

## Task: Object Tracking

* **Metrics:** MOTA (The Multiple Object Tracking Accuracy) [5]
* **Description:**

The MOTA accounts for all object configuration errors made by the tracker, false positives, misses (true negative), mismatches, over all frames.

where , , and are the number of false negative, the number of false positive, the number of mismatch error (ID Switching between 2 successive frames), and the number of objects in the ground truth respectively at time .

## Definition of Metrics ….

* **Metric 1:** mAP (mean Average Precision) [3] [4]
* **Task:** Object Detection / Object Segmentation
* **Definition:**

For a given category of object, true positive , false positive , false negative , and true negative are defined with an Intersection over Union (IoU) threshold for that category, where true/false represents the output of the neural network, positive/negative represents the label in the ground truth.

Then, recall of the given IoU threshold is defined as the proportion of all true positive examples in all true positive and false negative examples corresponding to that IoU threshold:

The precision of the given IoU threshold is the proportion of all true positive examples which are from all positive examples:

A neural network of detection or segmentation may achieve several pairs of recall and precision values corresponding to a certain IoU threshold and different confidence levels. For each recall value in the pairs, let takes the maximum precision value in all precision values for which the corresponding recall values are above the given recall value :

Average Precision (AP) of a given category of object is defined as the average value of for all recall values provided by the neural network, which can characterize the area of the entire precision-recall curve.

Mean Average Precision (mAP) is an averaged AP overall categories of objects and in a range of IoU thresholds. As an example, in MS COCO 2017 dataset, 10 IoU thresholds are taken at equal intervals from 0.50 to 0.95. In particular, AP50 and AP75 generally present the mAP when the IoU threshold is 0.50 and 0.75 respectively.

* **Metric 2:** MOTA (The Multiple Object Tracking Accuracy) [5]
* **Task:** Object Tracking
* **Definition:**

The MOTA accounts for all object configuration errors made by the tracker, false positives, misses (true negative), mismatches, over all frames.

where , , and are the number of false negative, the number of false positive, the number of mismatch error (ID Switching between 2 successive frames), and the number of objects in the ground truth respectively at time .

* **Metric 3:** bits per pixel (bpp)
* **Definition:**

bpp is the number of bits occupied by each pixel, which is defined by:

Total pixels refers to the total number of pixels over all images with original resolution.

Metric 4: PSNR (Peak Signal to Noise Ratio) [6]

Definition:

For two given images and of size , the PSNR of two images is defined by:

where is the dynamic range of the pixel values (255 for 8-bit grayscale images), and is defined by:

* **Metric 5:** SSIM (Structural Similarity) [7]
* **Definition:**

For two given signals and , the SSIM of the two signals is defined by:

where and are the average value of signals and respectively, and are the variances of signals and respectively, is the covariance of signals and , and are two constants. Specifically, we choose  and , where is the dynamic range of the pixel values (255 for 8-bit grayscale images), and are a small constant.