

**ISO/IEC JTC 1/SC 29/WG 11**

**Coding of moving pictures and audio**

**Convenorship: UNI (Italy)**

**ISO/IEC JTC 1/SC 29/WG 11 w19366**

**Document type: Approved WG 11 document**

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

**Status: Approved**

**Date of document: 2020-04-24**

**Source: Requirements**

**Expected action:**

**No. of pages: 7**

**Email of convenor: leonardo@chiariglione.org**

**Committee URL: mpeg.chiariglione.org**

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11/w19366**

**April 2020, Alpbach, AT**

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

**Source MPEG Requirements**

**Status:**

# 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 like VVC.
* Use the bitstream to support single or multiple tasks. Decompressed bitstream should be general enough to be usable 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.
* Allow the reconstruction of the compressed bitstream for human consumption. This can be achieved with an additional bitstream.

MPEG VCM has identified a set of relevant use cases and related requirements [1], focusing on the machine-to-machine communication in intelligent transportation and intelligent industry; and the hybrid machine and human consumption for surveillance and smart city use cases. This document contains information on how to provide evidence for these use cases. It contains details about:

* 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 needs 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-8.0 is used as the reference for the perfromance evaluation of MPEG-VCM encoder. Table 1 shows the tasks along with their metrics, datasets, benchmarks and traing/testing description.

At the current stage of this document, the key tasks for the MPEG-VCM perfomance evaluation are:

* Object Detection (still image)
* Object Segmentatioon (still image/video)
* Object Tracking (video)

*Table 1. Training/test conditions, key metrics, datasets, benchmark 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 | COCO [compressed]  (image) | [http://cocodataset.org/#detection-leaderboard](http://cocodataset.org/" \l "detection-leaderboard) | For COCO, use 2014 Val set for evaluation and 2014 Train in the case of retraining. |
| [CityScapes](https://www.cityscapes-dataset.com/) [uncompressed]  (image)  CityPersons  [uncompressed]  (image) | <https://www.cityscapes-dataset.com/benchmarks/> | For CityScapes, use defined train and validation sets |
| [ImageNet](http://www.image-net.org/) [compressed]  (image) | <https://kobiso.github.io/Computer-Vision-Leaderboard/imagenet.html> | For Imagenet, use the training and validation data as published from ILSVRC 2014. |
| Object Segmentation | [mAP](https://mc.ai/the-confusing-metrics-of-ap-and-map-for-object-detection/)  vs  BPP/Rate | COCO |  | (see above) |
| [CityScapes](https://www.cityscapes-dataset.com/) [uncompressed]  (image) |  | (see above) |
| [BD100K](https://bdd-data.berkeley.edu/)  (image) | <https://arxiv.org/abs/1805.04687> | We recommend using the predefined splits. |
| [KITTI](http://www.cvlibs.net/datasets/kitti/index.php)  (image) | <http://www.cvlibs.net/datasets/kitti/eval_object.php> | We recommend using the predefined splits. |
| [mAP](https://mc.ai/the-confusing-metrics-of-ap-and-map-for-object-detection/)  vs  Rate | DAVIS 2016 / 2017  (video) | <https://davischallenge.org/> | We recommend using the semi-supervised mode for higher accuracy. |
| Object Tracking | MOTA  vs  Rate | MOT20  [compressed]  (video) | <https://arxiv.org/pdf/1906.04567.pdf> | Dataset split is available from the Tracking Challenge, available on their website. |

* Datasets
  + Datasets must be of high quality such that a VVC encoder can create several coded versions of the sequence with noticeable degradation in quality
  + Datasets shall be capable of generating anchors per requirements
  + Proponents are invited to look into the datasets and raise a flag in case of encountering issues such as copy rights, etc. (Note: It is intended to have these datasets to be downloadable from the MPEG site)
* Reference codec configuration
  + VVC
  + Reference software VTM-8.0
  + 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
  + Performance:  mAP/MOTA vs BPP/rate
    - Specify max. endpoint (uncompressed) performance (mAP/MOTA)
    - Specify min. endpoint (threshold) performance (mAP/MOTA)
    - Checkpoints are uniformly distributed between max and min endpoints (including the end points)
    - Resolutions: 100%, 75%, 50% and 25%
    - Quantization Parameters (QP): 22, 27, 32, 37, 43, 48, 51
    - Generate performance (mAP/MOTA v.s. BPP) curves for each task based on the 4x resolutions and QP within the range of acceptable performance (min/max endpoints)
    - Performance curve is the Pareto front created from the 4 curves
* Network architectures
  + Object Detection: Faster R-CNN X101-FPN (part of Facebook AI Research’s Detectron)
  + Object Segmentation: R50-FPN Cityscapes (part of Facebook AI Research’s Detectron)
  + Object Tracking: JDE-1088x608
  + Optional requirement
    - As an additional performance data, other network architectures are allowed to be used per tasks. However, it is up to the proponents to provide a comparable performance data with the anchors specified in this document

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 unprocessed or 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. 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 comparison to the released compression ratio of VVC on the unprocessed video. For human consumption use cases, proponents shall report BD-rate. BD-rate should be calculated in the way as other standardization groups, e.g. JVET [2].

MPEG-VCM will release a stream of feature maps taken from video for several use cases based on current state of the art computer vision, along with the performance measured across different datasets. MPEG-VCM will also release the compression ratio and performance graphed for VVC with a common profile. Proponents are asked to test their compression coding on this stream, and shall report the bitrate of the compressed stream to the provided stream and report this ratio. Proponents are also asked to report the performance on the chosen database in the case of lossy coding.

# 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.

* Use case specific performance metrics, with the key tasks and metrics as defined above. Proponents shall perform the evaluation themselves, with the experiment 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 common neural network or general feature extraction methods regarding the specific key tasks. As an example, these common neural networks backbones may be VGG, ResNet, Inception and the specific frameworks depend on the key tasks. For detection and segmentation, an example may be Mask RCNN or YOLO.

# Evaluation Approach on Machine and Human Consumption

The evaluation process of video test data set for machine consumption and human consumption using VVC as an anchor are shown in Figure 1 and Figure 2, respectively.

|  |
| --- |
|  |

*Figure 1. Evaluation approach of video test data using VVC anchor - machine consumption*

|  |
| --- |
|  |

*Figure 2. Evaluation approach of video test data using VVC anchor - human consumption*

The feature data type and format information is beneficial for compression experts to know the properties and limitations about 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 2 shows an overview of different data types for different tasks required by the use cases described in [1].

*Table 2. 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 |

# References

1. w19365, Use cases and draft requirements for Video Coding for Machines, Alpbach, AT, October 2020.
2. K. Andersson, F. Bossen, J.-R. Ohm, A. Segall, R. Sjöberg, J. Ström, G. J. Sullivan. Summary information on BD-rate experiment evaluation practices. JVET-Q2016 (w19168), Brussels, Belgium, January 2020.