 **ISO/IEC JTC 1/SC 29/ WG 11 W19508**

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

**Coding of moving pictures and audio**

**Convenorship: Japan (JISC)**

|  |  |
| --- | --- |
| **Document type:** | Approved WG 11 document |
| **Title:** | **Draft Call for Evidence for Video Coding for Machines** |
| **Status:** | Approved |
| **Date of document:** | 2020-07-03 |
| **Source:** | Convenor, ISO/IEC JTC 1/SC 29/WG 11 |
| **No. of Pages:** | 8 |
| **Email of acting convenor** | ostermann@tnt.uni-hannover.de |
| **Committee URL:** | <http://isotc.iso.org/livelink/livelink/open/jtc1sc29> |
|  |  |

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2018/w19508**

**June-July 2020, Online**

**Title** **Draft Call for Evidence 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 a previously extracted feature stream and/or an optional video stream. 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. Features should be general enough to be used for different scenarios, for example, object detection and segmentation.
* Support varying performance levels for multiple tasks as measured by the appropriate metrics. This performance level may depend on the application.
* Allow the reconstruction of the original input for human consumption. This may be achieved with an additional bitstream.

MPEG VCM has identified a set of relevant use cases and related requirements [1] focusing on machine-to-machine communication and the hybrid machine and human consumption in intelligent transportation, intelligent industry, surveillance, smart city, and intelligent content 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

MPEG VCM is thus calling for evidence on compression technology on a processed or unprocessed video (ex. a stream of feature maps), and evidence of a shared set of features generated from methods such as neural network models for different tasks, which applies to the different use cases.

# Scope

The scope of this CfE is technology to reduce the size of the compressed processed or unprocessed video and evidence of a general feature extractor to enable a single task or multiple tasks while maintaining reasonable performance. The starting point for feature extraction is a shared feature extractor for carrying out single or multiple key tasks within a particular use case. On the other hand, the feature stream for a trained neural network is the starting point for the feature compression.

These use cases are defined in the Use Cases and Requirements Document [1].

Two specific technologies require evidence. The first technology is the compression of processed or unprocessed video. The main measure for feature compression is the compression ratio, while the complexity of compression and possible description of protocols will be taken into account. MPEG-VCM will provide data streams for the example use cases as listed in [1], and proponents are required to report the compression ratio of the compressed stream to the unprocessed video along with the performance of the decompressed stream.

The second technology is the shared backbone of feature generation to be used in a single task or multiple tasks. The performance assessment of the shared feature encoding is based upon the individual metrics as defined in [2]. The necessity of retraining and additional requirements will be taken into account.

Proponents may submit evidence for either technology or both technologies.

# Preliminary Timeline

|  |  |
| --- | --- |
| 2020-10-16 | Release of Call for Evidence document |
| 2020-10-16 | Availability of neural networks, test data, and description for the respective use cases. Pointers for the datasets for training and evaluation are described in [1]. |
| 2021-01-04 | Registration deadline |
| 2021-01-06 | Deadline for submission of descriptions (MPEG input contribution) of approaches and evaluation results (for both evaluation process of feature compression and a shared backbone) |
| 2021-01-09 | Evaluation of responses |
|  |  |

# Test Conditions

The test conditions, including metrics, datasets, and benchmarks are described in [2] and the Test Conditions Document.

# Evaluation Methods and Procedures

The evaluation procedure and metrics are described in the Evaluation Framework document [2]. The metrics consist of two parts, one relating to feature extraction and one relating to compression of processed or unprocessed video:

* Usecase specific performance metrics, with the key tasks and metrics as defined above. Proponents shall perform the evaluation themselves, with the experimental conditions described in [2].
* 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 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.

The results must be reported in an input document to MPEG-133, and the implementation details may be reported but are not required.

# Submission Requirements

The following material is to be submitted electronically. The material shall also be brought to the 133rd MPEG meeting.

The submission must contain but is not limited to:

* Comparison of the compression ratio of coded stream in comparison to the original video using VVC
* Performance results of the decoded bitstream for multiple tasks in a specific use case along with the performance results of state-of-the-art techniques. The performance should be measured by the metrics for different key tasks as defined in the Evaluation Framework document [2], such as for machine vision and human vision. Proponents should describe whether or not the task-specific network was modified from the original.

For the compression of processed or unprocessed video, additional information that the submission may contain is:

* preferably a description of the compression approach
* indication whether retraining has been performed, and a reference to the data set used for retraining
* preferably a description of the task-specific network

For the extracted features, additional information that the submission may contain is:

* preferably a description of the shared feature backbone
* indication whether joint training occurred or on which dataset the trained feature extractor weights were provided
* preferably a description of the task-specific network, and any modifications to the task-specific network

In the case of multiple tasks, any modification or retraining of the task-specific networks is allowed and should be reported.

# Participation fee

Participation in the call will not be associated with any fee.

# Logistics

Prospective contributors of responses to the Call for Evidence should contact the following people:

Jörn Ostermann (MPEG requirements chair)

Leibniz Universität Hannover.

Institut für Informationsverarbeitung

Tel. +49-5117625316, email [ostermann@tnt.uni-hannover.de](mailto:ostermann@tnt.uni-hannover.de)

M. Rafie

Gyrfalcon Technology Inc (GTI)

Tel. +1 4082059400, email [m.rafie@gyrfalcontech.com](mailto:m.rafie@gyrfalcontech.com)

Yuan Zhang

China Telecom

Tel. +86 18918588990, email [zhangy666@chinatelecom.cn](mailto:zhangy666@chinatelecom.cn)

* Expressions of interest to submit a response shall be made by contacting the people above on or before 2020-11-16. Interested parties are kindly invited to express their intent as early as possible.
* Details on how to format and submit documents, bitstreams, and other required data will be communicated directly to those who express an interest of participation.

Details for access to the test data and tools for evaluation can be found in [2], for further questions contact one of the above individuals.

# References

|  |  |
| --- | --- |
| [1] | w19506, "Use cases and draft requirements for Video Coding for Machines," Online, July 2020. |
| [2] | w19507, "Evaluation Framework for Video Coding for Machines," Online, July 2020. |