

**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 w18966**

**Document type: Approved WG 11 document**

**Title: Use cases and draft requirements for Video Coding for Machines**

**Status: Approved**

**Date of document: 2020-01-17**

**Source: Requirements**

**Expected action:**

**No. of pages: 9**

**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/w18966**

**January 2020, Brussels, Belgium**

**Title: Use cases and requirements for Video Coding for Machines**

**Source Requirements Subgroup**

**Status:**

# Introduction

## Motivation

Traditional coding methods aim for the best video/image under certain bit-rate constraint for human consumption. However, with the rise of machine learning applications, along with the abundance of sensors, many intelligent platforms have been implemented with massive data requirements including scenarios such as connected vehicles, video surveillance, and smart city.

The sheer quantity of data being produced constantly leads previous methods with a human in the pipeline to be inefficient, and unrealistic in terms of latency and scale. There are additional concerns in transmission and archive systems which require a more compact data representation and low latency solution. This led to the introduction of Video Coding for Machines.

In some cases, machines will communicate amongst themselves to perform tasks without a human in the mix, while in others there will be a need for additional human consumption of the specific decompressed stream. This specific scenario is possible in surveillance use cases, where a human “supervisor” may occasionally search for a specific person, or scene in video. In other cases, the corresponding bitstream may be used for both human and machine consumption. In the case of connected cars, the features may be used for image enhancement functionality for humans and object segmentation and detection for machines.

Any use cases in which video features need to be transmitted for additional processing which may potentially be used for machine or human end users could benefit from a standard in the coded features (shared backbone). Interoperability is crucial where different manufacturers and platforms need communication to achieve a common goal. Additionally, the feature stream must be efficient for both transmission and archive concerns for both latency and space. A standard for the compressed coding of this feature stream will establish an efficient protocol for machines to communicate.

## Scope

MPEG aims to define a compressed bitstream for extracted features in video for further processing on different nodes. This bitstream should be able to

* Enable efficient and high-performing solutions to multiple tasks
* Compares favorably in comparison to the original video file after compression with regards to performance and bitrate

## System Overview

The generic system architecture contains: a feature extractor that returns either processed or unprocessed video, and a compression scheme optimized for features. The feature extractor and compression scheme can be optimized for either a single task or multiple, and the size of the compressed stream should compare favorably to traditional coding techniques on the unprocessed video. The features may take different forms as described in below.

The decompressed output of the feature maps may then be used for post-processing tasks, which may include machine consumption tasks and human consumption tasks.

The output of the Feature Extraction Network is a standardized feature stream, which could include a stream of features, low bitrate video, and other information such as a list of key points. The decompressed bitstream can be used for both key tasks for machine vision and human vision. There is an optional profile from NNR for the Feature Extraction Network and the Task Specific Networks. The VCM decoder may be updated by sending appropriate components of the decoder, typically compressed using NNR. The MPEG activity on Video Coding for Machines (VCM) aims to standardize a bitstream format generated by compressing a previously extracted feature stream and an optional video stream.

Transmission storage

Task

Analysis for Machine vision

Human vision

Feature

De-compression

VCM decoder

Video

Feature Extraction

Network

Feature Compression Coding

VCM

Decompressed

bitsream

Feature

stream

Compressed

bitstream

NNR

NNR

Fig1. Pipeline for VCM

# Use Cases

|  |
| --- |
| UC1 Surveillance |
| Recently, surveillance systems have incorporated the use of neural networks for different tasks such as object detection and tracking. However, current surveillance systems often take up large amounts of data for storage due to the number of sensors and length of video to be recorded. |
| **Overlap with other use cases** |
| Subset of Smart City |
| **Required properties of the algorithm** |
| * It should be possible to enable object detection and instance segmentation, key points detection on the decompressed bitstream amongst other tasks * It should be feasible to reconstruct the video for human consumption. |
| **Optional properties of the algorithm** |
| * In some cases, the reconstructed image could also be enhanced for super resolution, or low light exposure. * In some cases, in order to protect privacy, the features should be sufficient for intelligence tasks but insufficient to reconstruct video when the additional video-reconstruction bit-stream is not used. * In some cases, the encoder should be capable to support multiple formats of input bitstreams. * In some cases, the decoder should be capable to detect falsified video * In some cases, it is needed to handle distorted or low quality video (ex. aerial recording, mobile vs non-mobile sensor) |
| **What are the different sub-tasks expected in this use case?** |
| * Object detection with object class, object Id, and object location * Instance Segmentation with object class, object Id, pixel-level object boundary or object shape. * Object Tracking with Object Id, Object appear or disappear indication. * Image Search * Image reconstruction * Image Enhancement   + Super resolution   + Low light enhancement * Event Recognition * Event Prediction * Anomaly Detection   + Malfunctions * Density Estimation   + Crowd density over a certain bounding box |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * Compared to other use cases, the required bandwidth might be larger since most surveillance systems are stationary. The likely limitation for bandwidth is the available storage. * The aerial surveillance system provides more flexible vision and continuous tracking of objects. The video resolution depends on wireless network connection speed. * The compressed feature stream may need to support real time or near real-time tasks * Image reconstruction/enhancement may not need to be in real time * Power is not a large concern |

|  |
| --- |
| UC2 Intelligent Transportation |
| In smart traffic system, cars may need to communicate features between each other and other sensors in order to perform different tasks. Sensors in the infrastructure may communicate features towards different vehicles, which then use these features to do object detection, lane tracking, etc. Final processing of these features is done on the individual vehicles. |
| **Overlap with other use cases** |
|  |
| **Required properties of the algorithm** |
| * It should be possible to do object detection and semantic segmentation on the decompressed features * It should be possible to reconstruct the bitstream for human consumption when needed. |
| **What are the different sub-tasks expected in this use case?** |
| * Object detection, with class Id, location, * Semantic Segmentation * Object Tracking * Image Enhancement * Event Recognition * Event Prediction |
| **What is the expected bandwidth of the distribution channel?**  **What is the typical latency that is acceptable?**  **What are the power requirements?** |
| * Bandwidth may be constrained to data networks (5G) * Applications need to be in real time, so latency must be low (5-10ms) * Power needs to be low in order to fit on transportation vehicles |

|  |
| --- |
| UC3 Smart City |
| With the rise of IoT, there is a high degree of interconnectivity between different node sensors and devices. It is important for these devices to communicate with each other to optimize and efficiently solve tasks. Different vendors may develop part of the VCM pipeline, and there is a need for interoperability between devices and systems. Smart City applications encompass use cases such as traffic monitoring, density detection and prediction, traffic flow prediction and resource allocation. |
| **Overlap with other use cases** |
| Superset of surveillance use case |
| **Required properties of the algorithm** |
| * It should be possible for object detection and instance segmentation, key points detection on the decompressed bitstream * It should be feasible to reconstruct the video for human consumption. |
| **Optional properties of the algorithm** |
| * Optionally, the reconstructed image could also be enhanced for super resolution, or low light exposure. |
| **What are the different sub-tasks expected in this use case?** |
| * Object detection with object class, object Id, and object location * Instance Segmentation with object class, object Id, pixel-level object boundary or object shape. * Object Tracking with Object Id, Object appear or disappear indication. * Event Detection and Prediction * Image Reconstruction * Image Search * Image Enhancement   + Super resolution   + Low light enhancement |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * There may be larger bandwidth compared with other cases because most surveillance systems are non-mobile. The limitation with the bandwidth most likely comes from the storage requested. * Image reconstruction/enhancement may not need to be in real time * The generation of the feature stream may need to be done in real-time or near real-time * Power is not a large concern |

# Summary of Proposed Sub-tasks

|  |  |  |  |
| --- | --- | --- | --- |
|  | Description | Surveillance / Smart City | Intelligent Transportation |
| Object Detection | Determine a bounding box for an object that may be in the input image / video along with object id | x | x |
| Object Segmentation | Determine which pixels belong to which objects by defining binary masks for each image | x | x |
| Image/Video Reconstruction | Given the compressed feature stream with an additional bit-stream return the reconstructed image/video | x |  |
| Image/Video Enhancement | With an additional bit-stream return the reconstructed image/video enhanced for human consumption such as super resolution, low light | x |  |
| Object Tracking | Determine the location of an object throughout video along with object id | x | x |
| Event Recognition | Determine which event has occurred in the video | x | x |
| Event Prediction | Predict which event will occur | x | x |
| Anomaly Detection | Determine whether or not a nonstandard deviation has occurred such as malfunctions | x | x |
| Density Estimation | Estimation of population density within a certain bounding box | x |  |
| Event Search | Provide a time stamp for when an event has occurred given an input image or video | x |  |

# Metrics for Key Tasks

Different datasets and metrics for key tasks are defined in the evaluation framework document.

# Requirements

The term “machine” refers to a process or algorithm that gets as input video data (eventually after a decoding stage) in order to analyse it or process it. For example, a machine is a neural network with the task to detect people in the input video.

|  |  |
| --- | --- |
| **Requirement** | **Description** |
| Efficient compression of bitstream | The size needed to represent the coded features shall be less than the encoded video stream under traditional video coding techniques with comparable performance at reasonable setting choices  Image size can vary depending on dataset and model input sizes. The compressed bitstream for still images should be smaller than JPEG2000 or other state of the art. For video, the compressed bitstream should be more compact than HEVC at similar performance levels. |
| Ability to use bitstream to support both single and multiple tasks | The resulting coded features shall be usable and optimized for different scenarios   * Only one specific machine [architecture, training and task]. Example: ResNet ImageNet classifier * Only one machine type [task]. Example: scene-level classification. * Multiple machine types. Examples: object-level classification AND action recognition.   There could be an advantage for using the bitstream for single tasks in comparison to using the bitstream for additional tasks. The bitstream for single tasks may either be smaller or the encoder complexity may be lower than the encoder for additional tasks. |
| Varying degrees of performance for multiple tasks | Some machines may be required to perform more accurately than others (i.e., the tasks that some machines perform may have higher priority or importance than the tasks performed by other machines).  Priority may be a function of latency, bandwidth, or other application-specific requirements which may result in the varying encoding of the video stream.  The coding shall support varying levels of quality as measured by performance for different sub tasks. |
| Reconstructable bitstream for human inspection | The decompressed bitstream should be suitable for human consumption. This may require an additional bitstream.  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 HEVC encoding of the unprocessed video. |
| Configurable profile for feature extraction network and task-specific network | The interface for the feature extraction network and task-specific networks should be configurable according to a profile as defined in MPEG-NNR. |

# Additional Use Cases

## 6.1 Machine vision use case list

Table 1 List of description and compression of features for machine analysis use cases

|  |  |  |  |
| --- | --- | --- | --- |
| item | Machine-oriented Analysis Use Cases | Description | Techniques |
| 1 | Smart Glasses | Record daily activity log,  Navigation (indoor/outdoor w/o GPS),  Video recording wake up | Object detection, Segmentation |
| 2 | Unmanned store | tracking customer activity,  check items in shopping cart | Object detection, Pose estimation, Object tracking |
| 3 | Unmanned Warehouse/Store Robot | Robot navigation, stocking, inventory checking | Detection, Segmentation, Classification |
| 4 | Smart Retailer | * Shopping Center Customer Group Analysis, * Detect hot spot inside store by customer age, gender. * Customer Traffic information | Detection,  Heat map,  Activity analysis. |
| 5 | Industrial Production Line Detect Equipment | Detect failure elements in production line for quality control | Detection,  Tracking |
| 6 | Smart factory /Automatic Machinery | Detect appearance of object, pick up and transfer to destination | Detection,  Segmentation,  Key point detection |
| 7 | Smart fishery/agriculture | Detect diseases | Detection,  Classification |
| 8 | Drone | Real time environment monitoring and automatic collision avoidance | Detection,  Segmentation,  tracking |

## 6.2 Human and Machine vision Use case list

Table 2 List of combined human/machine-oriented video representation and compression use cases

|  |  |  |  |
| --- | --- | --- | --- |
| item | Combined Machine and Human representation Use Cases | Description | Techniques |
| 1 | AR/VR and Video Game Goggles | Capture live video and detect environment elements | Detection, Segmentation,  Pose Estimation, Tracking |
| 2 | Sports Game animation | From live game video, create animation | Detection,  Segmentation,  tracking |

# References

[1] Compact Descriptors for Video Analysis: Objectives, Applications and Use Cases, ISO/IEC JTC1/SC29/WG11/W14507, 2014, Valencia.

[2] Use Scenarios of CDVA for Surveillance Domain, ISO/IEC JTC1/SC29/WG11/N15043, October 2014, Strasbourg, FR.