 ISO/IEC JTC 1/SC 29/WG 2 N43

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

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

**Document type:** Output Document

**Title: Use cases and requirements 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:** 19 (without 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 N43**

**Online – January 2021**

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

**Source: WG 2 MPEG Technical requirements**

**Status: Approved**

**Serial number: 20127**

**Editor(s): Yuan Zhang(China Telecom), M.Rafie(GTI), Shan Liu(Tencent)**

# Introduction

## Motivation

Traditional coding methods aim for the best video 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-VCM aims to define a bitstream from compressing video or feature extracted from video that is efficient in terms of bitrate/size and can be used by a network of machines after decompression to perform multiple tasks without significantly degrading task performance. The decoded video or feature can be used for machine consumption or hybrid machine and human consumption.

The differences between VCM and video coding with deep learning are:

1. VCM is used for machine consumption or hybrid machine and human consumption, while current video coding aims for human consumption;
2. VCM technologies could be but is not required to be based on deep learning;
3. VCM can achieve analysis efficiency, computational offloading and privacy protection as well as compression efficiency, while traditional video coding pursues mainly on compression efficiency.

## System Overview

The generic system architecture contains a pair of VCM encoder and decoder. The input of the VCM system could be video or features. In case of a feature stream, the type and format of feature should be specified, features may take different forms as described in Evaluation Framework for Video Coding for Machines document[1].

The decompressed bitstream of video and/or feature may then be used for post-processing tasks, which may include machine consumption tasks or hybrid machine and human consumption tasks. The encoder 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 MPEG activity on Video Coding for Machines (VCM) aims to standardize a bitstream format generated by compressing a previously extracted feature stream or video stream.

VCM

Encoder

VCM

Decoder

Compressed

bitstream

Task Analysis for

Machine Vision

Human Vision

(optional)

（optional）

Decompressed

Video/Feature

Video/Feature

Fig1. Pipeline for VCM

Transmission of encoded bitstream

VCM decoder

VCM encoder

Feature Extraction

Feature Encoding

Feature Decoding

Interface for NN

Interface for NN

Human Vision

Machine Vision

Sensor output

Video Encoding

Bitstream Mux

Bitstream Demux

Video Decoding

Feature Conversion

Fig 2. An example of potential VCM architecture

Fig 2 shows an example of potential VCM architecture. The VCM codec could be video codec or feature codec, or both. In case of feature codec, the VCM feature encoding is consisted of feature extraction, feature conversion and feature coding. There may be an interface to NNR for the feature extraction and the task specific networks. Regarding the detailed examples of VCM pipelines, please refer to the Evaluation Framework of Video Coding for Machines document [1].

# 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, non-visible light images such as IR images and LIDAR are used for low-light and atmospheric conditions. * 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 * Stereoscopic and multiview video processing * Pose estimation * Pose tracking * Action Recognition * 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. |
| **Optional properties of the algorithm** |
| * In some cases, non-visible light images such as IR images and LIDAR are used for low-light and atmospheric conditions. |
| **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 * Stereoscopic and multiview video processing * Pose estimation * Pose tracking * Action Recognition |
| **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. * In some cases, non-visible light images such as IR images and LIDAR are used for low-light and atmospheric conditions. |
| **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 * Stereoscopic and multiview video processing * Pose estimation * Pose tracking * Action Recognition |
| **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 use 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 |

|  |
| --- |
| UC4 Intelligent industry |
| With the rapid development of intelligent industry, the degree of automation production has been enhanced, making working environments which is unsuitable for manual work possible, and large-scale, continuous production a reality. Production efficiency and accuracy are greatly improved. Different vendors may be part of the VCM pipeline, and interoperability is required for devices to post-process the features to perform multiple tasks. |
| **Overlap with other use cases** |
|  |
| **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 low light exposure. * In some cases, it is needed to handle distorted or low quality video (i.e., mobile device). * In some cases, non-visible light images such as IR images and LIDAR are used for low-light and atmospheric conditions. |
| **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 * Low light enhancement * Measurement of object parameters (size, orientation, curvature, angle) * Event recognition * Anomaly Detection * Video masking/concealment * Stereoscopic and multiview video processing * Pose estimation * Pose tracking * Action Recognition |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * In some cases, bandwidth may be constrained to data networks (5G). * The compressed feature stream may need to support real time or near real-time tasks. * Power is not a large concern. |

|  |
| --- |
| UC5 Intelligent Content |
| Due to the generation of huge number of video/image contents, various forms of media including conventional broadcasting and personal broadcasting are overflowing, to protect certain groups of people (i.e., people under 18) from inappropriate content is becoming a big issue. Meanwhile, the traditional manual review is time-consuming and labor-intensive. Machine vision technologies help live images/videos, short videos, and social media perform intelligent review, rating, processing, and distribution. |
| **Overlap with other use cases** |
|  |
| **Required properties of the algorithm** |
| * It should 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 low light exposure. * In some cases, it is needed to handle distorted or low quality video (i.e., mobile device). |
| **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   + Event detection, recognition and prediction   + Image search   + Image reconstruction   + Image enhancement   + Stereoscopic and multiview video processing   + Pose estimation   + Pose tracking   + Action Recognition   + Anomaly detection |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * In some cases, bandwidth may be constrained to data networks (5G). * In some cases, the compressed feature stream may need to support real time or near real-time tasks. * Power is not a large concern. |

|  |
| --- |
| ***2.6 UC6 Consumer electronics*** |
| Consumer electronics use neural networks to provide context-aware services and information to users. It is important to communicate features with each other to obtain information to perform tasks. It can also be transmitted to an external server such as a cloud or edge server. However, when communicating using traditional video, the required network bandwidth can be burdened and privacy may be compromised.  VCM can also provide interoperability between devices from different vendors. Low bandwidth communication and personal information protection are also desirable. |
| **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. |
| **Optional properties of the algorithm** |
| * 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. |
| **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   + Event detection, recognition and prediction   + Image search   + Image reconstruction   + Image enhancement   + Pose estimation   + Pose tracking   + Action Recognition   + Anomaly detection   + Stereoscopic and multiview video processing |
| **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?**  **What are the power requirements?** |
| * In some cases, bandwidth may be constrained to data networks (4G/5G). * In some cases, the compressed feature stream may need to support real time or near real-time tasks. * Power and computational complexity could be of concern in some cases, especially in battery operated devices. |

# Summary of Proposed Sub-tasks

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Description | Surveillance / Smart City | Intelligent Transportation | Intelligent Industry | Intelligent Content | Consumer electronics |
| Object Detection | Determine a bounding box for an object that may be in the input image / video along with object id | x | x | x | x | x |
| Object Segmentation | Determine which pixels belong to which objects by defining binary masks for each image | x | x | x | x | 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 |  |  | x | x |
| Object Tracking | Determine the location of an object throughout video along with object id | x | x | x |  |  |
| Event Recognition | Determine which event has occurred in the video | x | x | x | x | x |
| Event Prediction | Predict which event will occur | x | x |  | x | x |
| Anomaly Detection | Determine whether or not a nonstandard deviation has occurred such as malfunctions | x | x | x | 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 |  |  | x |  |
| Measurement | Measure the object parameters (size, orientation, curvature, angle) |  |  | x |  |  |
| Object masking | Detect and conceal the certain object in video with a mask | x |  |  | x | x |
| Stereoscopic and multiview video processing | Depth estimation , view synthesis, and multiview video coding | x | x | x |  | x |
| Pose estimation | Skeleton information estimation in a frame for objects | x | x | x | x | x |
| Pose tracking | Skeleton information estimation in every frame in a video with ID for objects in the video | x | x | x | x | x |
| Action Recognition | Action recognition in the video | x | x | x | x | x |
| Image/Video Reconstruction category 1 | Reconstruct the image/video  from the compressed feature stream  for human consumption of entertainment |  |  |  | x | x |
| Image/Video Reconstruction category 2 | Reconstruct the image/video  from the compressed feature stream  for human consumption of the same tasks as machine vision | x | x | x |  |  |
| Image/Video Reconstruction category 3 | Reconstruct the image/video  from the compressed feature stream  for human consumption of tasks that machine-vision not target to | x | x | x |  |  |

# Metrics for Key Tasks

Different datasets and metrics for key tasks are defined in the evaluation framework document[1].

# Requirements

The term “machine” refers to a process or algorithm that gets as input video or feature (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.

The potential benefits of VCM include compression efficiency, computational offloading, and privacy protection, etc. The following requirements reflect different usecases. Requirements 1, 2 and 3 are required. Requirement 4 is optional. Requirements 5, 6 are recommended.

|  |  |  |
| --- | --- | --- |
| **Number** | **Requirement** | **Description** |
| 1 | Efficiency of compression of bitstream | The size needed to represent the video or 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. |
| 2 | Coding technologies have the ability to generate one bitstream to support single task or multiple tasks | The resulting coded features shall be usable and optimized for different scenarios   * One bitstream for single task. Example: scene-level classification. * One common bitstream for multiple tasks. Examples: object-level classification and action recognition.   The coding technology should support multiple tasks. 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. |
| 3 | 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. |
| 4 | Hybrid machine and human consumption | A common bitstream should be used for machine and human consumption, additional bitstream(s) is optional, in which human consumption of reconstructed video may include 3 categories:  (1) human consumption for entertainment;  (2) human consumption for the same tasks as machine vision; and  (3) human consumption for tasks that machine-vision not target to.  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. |
| 5 | Computational offloading | Due to the limited computational power of front-end devices or cloud servers, some machines are required to perform computational offloading: Part of the machine analysis tasks could be done on front-end devices, and outputs of intermediate layers of neural networks can be compressed and transferred to the cloud servers. The computational cost includes both machine analysis tasks and the compression. |
| 6 | Privacy protection | The features should be:  Not usable for reconstruction of image/video, or  Detected and processed according to their pre-defined privacy level  To ensure the inability to reconstruct the part of image/video related to privacy. |

# 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 | Unmanned store | tracking customer activity,  check items in shopping cart | Object detection, Pose estimation, Object tracking, Stereoscopic and multiview video processing |
| 2 | Unmanned Warehouse/Store Robot | Robot navigation, stocking, inventory checking | Detection, Segmentation, Classification, Stereoscopic and multiview video processing |
| 3 | Smart Retailer | * Shopping Center Customer Group Analysis, * Detect hot spot inside store by customer age, gender. * Customer Traffic information | Detection,  Heat map,  Activity analysis, Stereoscopic and multiview video processing |
| 4 | Smart fishery/agriculture | Detect diseases | Detection,Classification |
| 5 | UAV | Real time environment monitoring and automatic collision avoidance | Detection,  Segmentation,  tracking, Stereoscopic and multiview video processing |

## 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, Stereoscopic and multiview video processing |
| 2 | Sports Game animation | From live game video, create animation | Detection,  Segmentation,  tracking, Stereoscopic and multiview video processing |
| 3 | Smart Glasses | Record daily activity log,  Navigation (indoor/outdoor w/o GPS),  Video recording wake up | Object detection, Segmentation, Stereoscopic and multiview video processing |

# References

1. wxxxxx, Evaluation Framework for Video Coding for Machines, Online, January 2021.
2. w19841, Use cases and draft requirements for Video Coding for Machines, Online October 2020.
3. m55987 [VCM] Comments on VCM requirements, Online, January 2021.

# Annex Examples

# *A.1 Video concealment*

These days CCTVs are widely used to prevent crimes, or provide the useful information such as regional or traffic information. However, personal privacy infringement has been a big issue in the security field. In particular, it is frequently requested to prevent abuse of personal information in public and drone images.

In this example, to solve personal privacy problems in CCTV, VCM encoder detects privacy area in each scene, and then encodes and transmits the original video after masking the privacy area with VCM features. At the decoder, the privacy information can be restored only when the VCM feature is utilized. In this case, the legacy video decoder can only display the privacy masked video.

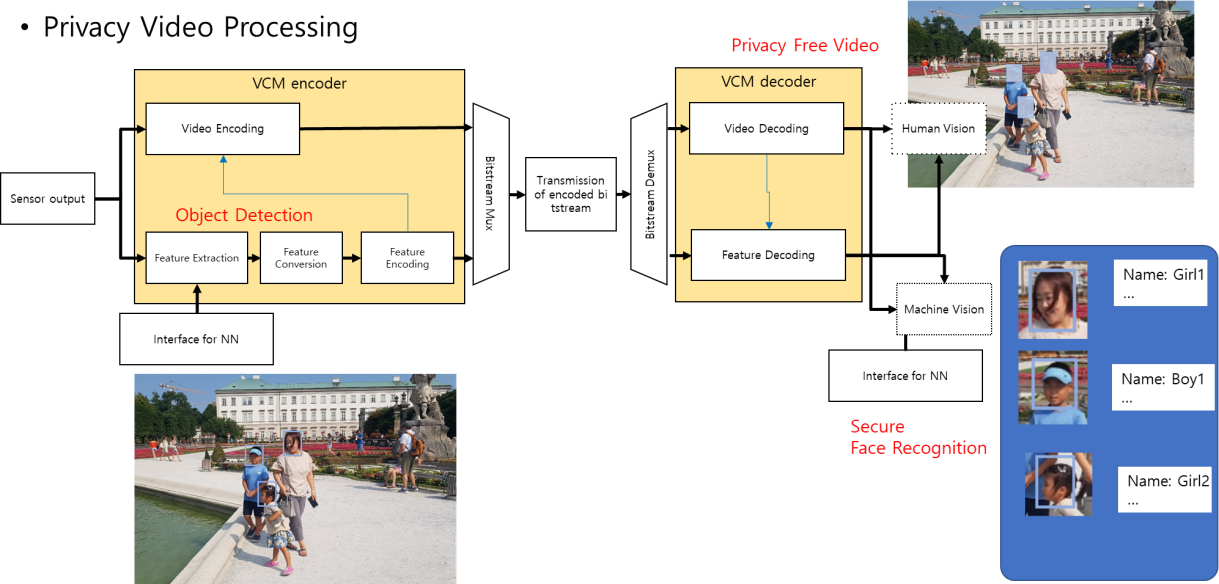


Fig 3. Privacy control in video surveillance

# *A.2 Content detection*

Another example is video analysis task such as explicit content detection or adult content detection. As various forms of media including conventional broadcasting and personal broadcasting are overflowing, to prevent child from inappropriate content are becoming a big issue.

Here, we consider a video service that can prevent these harmful contents depending on ages. In this example, VCM encoder automatically analyzes and detects hazardous objects and scenes from video using deep learning network. Then, it encodes video and features (i.e. detected hazardous items) and simultaneously transfers them to viewers. A viewer on VCM decoder side enters the profile information such as the user's age or goes through the real-time age verification step. And then, the video and features are conditionally decoded and reproduced to the scene according to user profiles. For example, video with masked hazardous objects and scenes is reproduced for teenagers or young boys conversely, video with highlighted hazardous items in bold line is reproduced only for identified adult viewers.

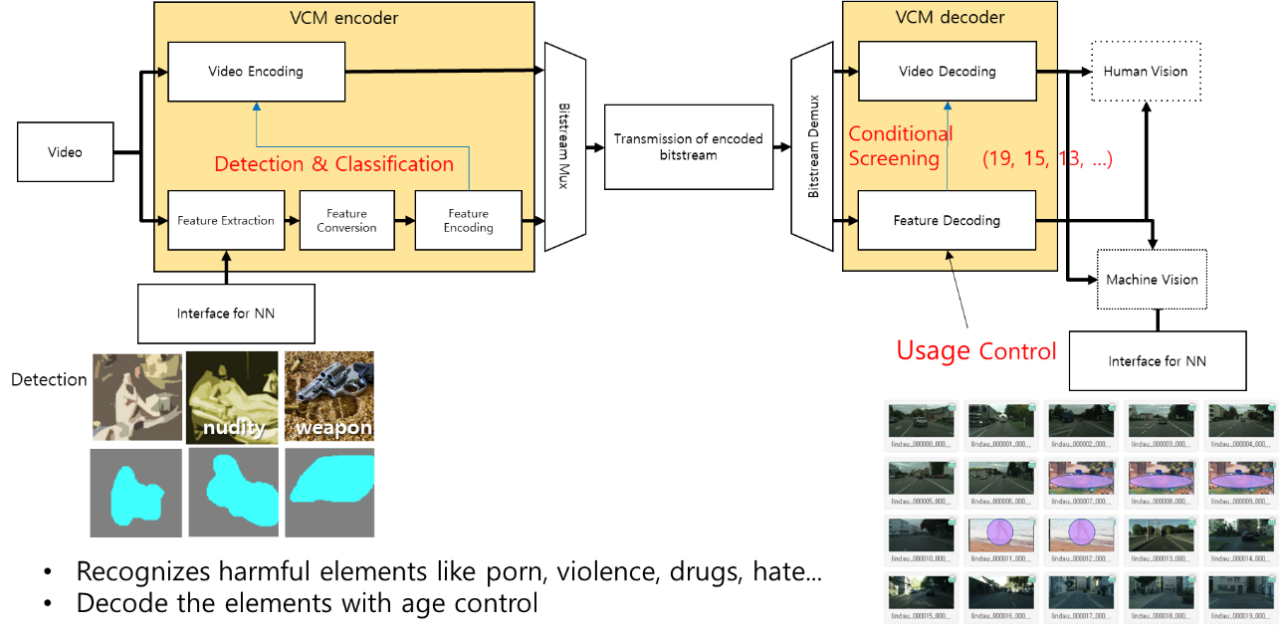


Fig 4. Content protection

## A.3 Feature from additional input channel

Transmission of features extracted from additional input channel such as invisible light is also considerable as another example scenario. Detection of invisible or hidden objects under night time or fog, rainy conditions is one of the important tasks for safety in autonomous driving. It is known that the use of multi-modalities such as RGB, IR, and LIDAR has the advantage of obtaining consistently reliable results over the use of single-modality under the various conditions. Furthermore, emerging ICT technologies enable communications among vehicles and infrastructures in the mobile environment of up to 250km/h. As the amount of sensor data and connectivity increases, we can expect that the need for VCM technology will also increase.

In this example scenario, the information of detected objects with original video is encoded by VCM encoder and the bitstream is transmitted via emerging telecommunication technologies for vehicles such as V2X. The connected vehicles and infrastructures can recognize hidden or invisible objects by decoding the received data, and use the data for machine-oriented tasks (such as path planning) and/or human-oriented tasks overlaid detection results on the video for assisted driving).

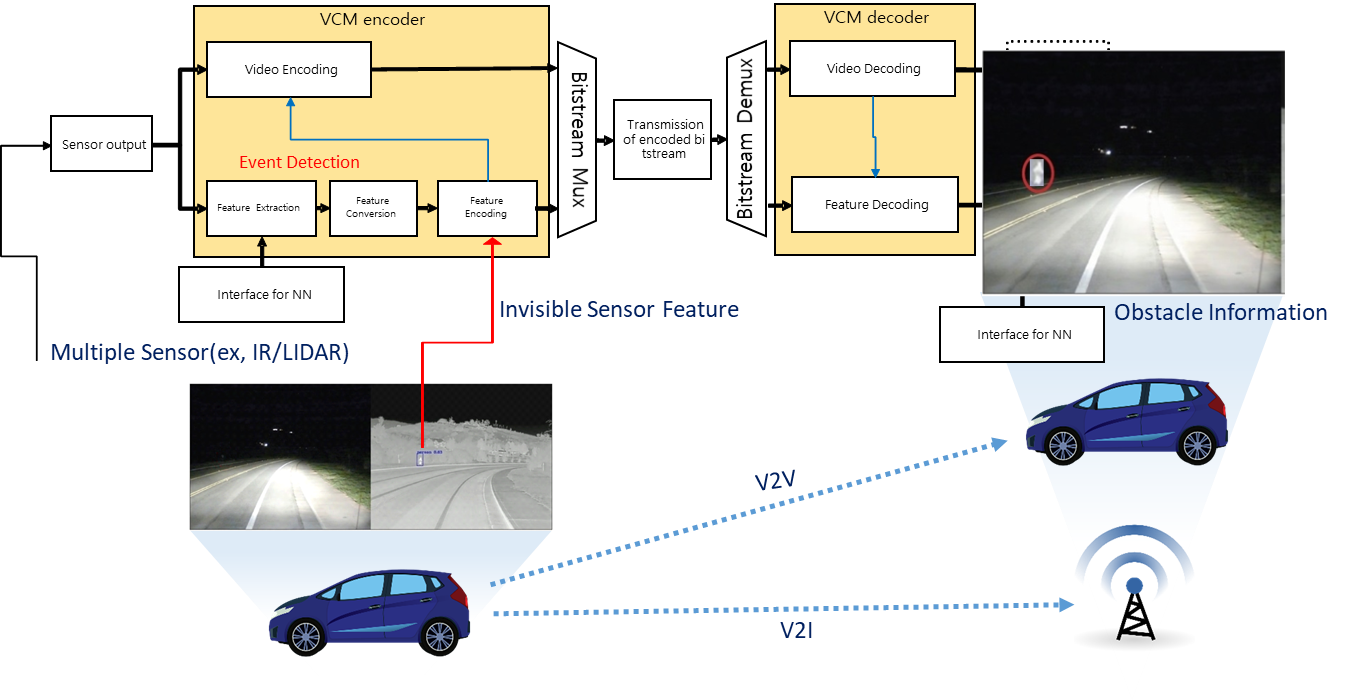


Fig 5. Hidden/invisible object alarm