

ISO/IEC JTC 1/SC 29 "Coding of audio, picture, multimedia and hypermedia information"

Secretariat: JISC

Committee manager: Koike Mayumi Ms.



Liaison statement from SC 29/WG 2 to 3GPP SA 1 on Video Coding for Machines and Haptic Use Cases [SC 29/WG 2 N 273]

Document type	Related content	Document date	Expected action
Project / Other		2023-02-01	INFO

Description

In accordance with Recommendation 4.3.1 at the 10th WG 2 Meeting, 2023-01-16/21, Virtual, the SC 29 Secretariat sends this liaison statement to 3GPP SA 1. [Requested action: For SC 29's information]

ISO/IEC JTC 1/SC 29/WG 2
MPEG Technical requirements
Convenorship: SFS (Finland)

Document type:	Output Document
Title:	Liaison to 3GPP/SA1 on Video Coding for Machines and Haptics Use Cases (CC SA4)
Status:	Approved
Date of document:	2023-01-20
Source:	ISO/IEC JTC 1/SC 29/WG 2
Expected action:	None
Action due date:	None
No. of pages:	3 (with cover page) + 17+31 (attachments)
Email of Convenor:	igor.curcio@nokia.com
Committee URL:	https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-2

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 2
MPEG TECHNICAL REQUIREMENTS

ISO/IEC JTC 1/SC 29/WG 2 N00273

Online - January 2023

Title	Liaison to 3GPP/SA1 on Video Coding for Machines and Haptics Use Cases (CC SA4)
Source	WG 2, MPEG Technical Requirements
Status	Approved
Serial Number	55072

ISO/IEC JTC 1/SC 29/WG 2 would like to inform 3GPP TSG SA WG1 about the following activities that may be of relevance in the course of your ongoing Release 19 Study on Network of Service Robots with Ambient Intelligence (FS_SOBOT) and any subsequent activity related to tactile communications:

MPEG-I Phase 2 Haptics

WG 2 has developed a series of use cases and requirements for haptics service scenarios. Each use case is documented with a technical description of the Haptics schema, required features and, certainly more relevant for 3GPP, potential haptics requirements and specifications. A wide range of applications are considered, and we would like to bring to your attention to the Telerobotic surgery use case. It contains end-to-end characteristics that may benefit from some functional support at the network level. The report is provided as attachment in ***wg2n00139***.

Video Coding for Machines (VCM)

Traditional coding methods aim for the best video under certain bit-rate constraints 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 cities. Video Coding for Machines is a process of encoding and decoding video, descriptor or feature extracted from video for a machine task. WG2 has developed a use cases and requirements document that also includes various architectures envisaged for the VCM codec and for communicating between end-points (e.g., UE to UE or UE to Server).

Use cases illustrating the need for a machine-type communication specific format and its characteristics are listed in the attachment ***wg2n00190***.

ACTION: WG2 asks 3GPP SA1 to take the above information into consideration and provide feedback if deemed necessary.

Dates of Next WG2 Meetings:

WG2#11	24 – 28 April 2023	Antalya
WG2#12	17 – 21 July 2023	Geneva

ISO/IEC JTC 1/SC 29/WG 2
MPEG Technical requirements
Convenorship: SFS (Finland)

Document type:	Output Document
Title:	Updated MPEG-I Phase 2 Haptics Use Cases
Status:	Approved
Date of document:	2021-10-15
Source:	ISO/IEC JTC 1/SC 29/WG 2
Expected action:	None
Action due date:	None
No. of pages:	31 (with cover page)
Email of Convenor:	igor.curcio@nokia.com
Committee URL:	https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-2

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 2
MPEG TECHNICAL REQUIREMENTS**

ISO/IEC JTC 1/SC 29/WG 2 N00139
Online – October 2021

Title	Updated MPEG-I Phase2 Haptics Use Cases
Source	WG 2, MPEG Technical requirements
Status	Approved
Serial Number	20966

Contents

1.	Introduction	2
1.1.	<i>Phasing of MPEG-I haptics</i>	2
1.2.	<i>Considerations for Haptic Specifications</i>	2
2.	Phase 2A Use Cases	3
2.1.	<i>Combined Point Cloud and Video 6DoF Contents</i>	3
2.2.	<i>Full Immersive Content: 6DoF with full 3D 360 video</i>	5
2.3.	<i>A Mobile subject looks around a statue Object with up-close views</i>	6
2.4.	<i>Entertainment/Immersive Video/Theatrical</i>	7
3.	Phase 2B Use Cases	9
3.1.	<i>VR Video Calling</i>	9
3.2.	<i>Virtual Prototyping</i>	12
3.3.	<i>Training and Simulation</i>	14
3.4.	<i>Telerobotic surgery</i>	15
3.5.	<i>Marketing/Sales</i>	17
3.6.	<i>Gaming/E-Sports</i>	18
3.7.	<i>Telepresence</i>	20
3.8.	<i>Driver Augmentation</i>	22
3.9.	<i>ATSC 3.0 Broadcasting</i>	23
3.10.	<i>Apparel Fitting Rooms/Retail</i>	24
	References	25
	Annex A: Haptics in Immersive Media – A Tutorial	1
	<i>Introduction</i>	1
	<i>Value of Immersive Content with Haptics</i>	1
	<i>Typology of Haptic Immersive Content</i>	1
	<i>Story-driven Content</i>	1
		1

Experience-driven Content	2
Event-driven Content	2
Interaction-driven Content	3
Content Consumption Modalities	3
Finger Gesture	3
Hand Gesture	4
Head Gesture	4
Dimensions of Haptic Design	4
Haptic Playback for Immersive Media	5

1. Introduction

This document is best considered as a more well-informed update of **N19513** (MPEG-I Phase 2 Haptics Use Cases, [1]) that lists the haptics use cases for immersive media. The use cases in this document reflect the latest thinking of the Haptics AHG as it has deliberated over the appropriateness and technical feasibility of use cases to be included in the Haptics Phase 2 effort. To that end, several of the use cases in N19513 have been either removed or merged and several new use cases, reflective of the current, increasing use of haptics in several fields, have been added. **Given the significant amount of new prose, only the prose carried over from N19513 is in red, to improve readability. The rest of the prose is in this document, in black, is new.** Further, Phase 2 has been split up into two sub-phases, Phase 2A and Phase 2B. The rationale for this split is explained briefly below.

For some more background on the use of haptics in immersive media and the various design approaches, see Annex A “Haptics in Immersive Media – A Tutorial”.

1.1. Phasing of MPEG-I haptics

The Phase 1 Haptic CfP proposes a coded representation for haptic media that will satisfy many non-interactive and interactive use cases. For the purpose of generating requirements, it is useful to consider a natural division of use cases for Phase 2 based primarily on the type of interactivity that they require:

Phase 2A: Use cases where haptics are distributed in space and associated with a scene description. Visual, audio, and haptic rendering are synchronized with the scene presentation.

Phase 2B: Use cases where haptic feedback is rendered as a result of interaction between the extended user avatar and an extended reality environment.

In the subsequent sections, each use case includes a ‘haptic schema’ section that describes a possible embodiment of the haptic functionality for the use case.

1.2. Considerations for Haptic Specifications

There are a number of key considerations related to haptic specifications that are discussed in more detail in (WG02 N00115 [2]). They are summarized here for convenience.

Consideration	Summary
No Reference Device	There is no single device that can generate the entire range of perceivable tactile sensations.
Perceptual Variance	Tactile sensitivity varies greatly over the body surface and by tactile modality and due to cross-modal effects (e.g., priming and masking).

Avatar vs Physical Body	The 2m ² skin surface (+ proprioception) is not usually well represented in virtual environments.
Physical Device Mapping	A single user may have multiple different tactile stimulation devices with different performance capabilities, at different body locations.
Closed vs. Open Loop Feedback	Certain types of feedback (e.g., kinesthetic) require low sensor-to-actuator latency closed-loop feedback with low-jitter refresh rates.
Synchronization with other Media	Haptic perception is strongly influenced by intermodal effects from visual and audio stimulation.
Interactivity Models	Haptics in XR implies an interaction with the virtual environment. This type of interactivity can be more involved than visual/audio-based interaction feedback because feedback is often based on direct virtual contact.

2. Phase 2A Use Cases

The following use cases are based on selected use cases from N19513 that have been deemed relevant to Phase 2A, based on the criteria listed in the Introduction above. The requirements at the end of each use case have been modified to match the requirements in [3].

2.1. *Combined Point Cloud and Video 6DoF Contents*

Description

A user is watching a sports match, or a concert, using a device with the capability to provide input from the user to enable him/her to change his/her viewpoint location and direction within the sport or concert venue, without restriction.

For example, the user has the possibility to select a viewpoint from a 1st person perspective of a sports player, or a 3rd person perspective viewpoint similar to that of more traditional TV broadcast sports contents (Figure 1).



Figure 1: An example of 2 different viewpoints which could be selected by a user. Left: 1st person player view. Right: 3rd person commentary view.

The 6DoF content which the user is viewing is rendered using a combination of both point cloud media data, and video media data. The whole sports or concert venue is captured using multiple high-resolution cameras, such that the video data captured can be processed to create a point cloud scene of the center of the sports or concert venue (e.g., the sports pitch, players and other dynamic objects are represented by point clouds in the scene). This processing can be performed either at the venue itself, or remotely on a dedicated network. The venue may also be equipped with sensors that can capture Tactile Essence (SMPTE st2100-1-2017 (Coding of Tactile Essence)).

By creating such point cloud media data, a user has the freedom to navigate within the sports or concert venue (i.e., the defined scene boundary here) and is able to view different players and objects from all viewpoints and positions.

Since such venues are traditionally very big, and include massive crowds of spectators, it is possible to represent such non-interactive parts of the scene background using video media data.

The result is that the user views both point cloud and video media rendered at the same time in order to create an immersive experience.

Haptic schema

- 1. Haptics associated with point cloud media – A haptic track may be associated with a set of point cloud media data.**
 - a. Subsets of the point cloud may be indexed against specific objects/participants.**
 - i. This information can be used to associate haptics with specific subsets of the point cloud.**
 - ii. Alternatively, each discrete subset of the point cloud could be encoded as a separate media stream with its own haptic track.**
 - b. This haptic track can be activated, deactivated, or modulated based on the user’s viewpoint.**
 - i. Activated/deactivated - If the point cloud media is within the user’s field of view, the haptic track can be activated, and if the point cloud media moves out of the user’s field of view, the haptics can be deactivated.**
 - ii. Modulated – As the point cloud approaches the center of the user’s field of view, and/or as it approaches the 3D coordinates of the user’s perspective, the haptic track can be increased in magnitude (volume). As the point cloud moves away, it can be decreased in magnitude.**
 - c. When transitioning from one haptic track to another based on shifts in visual perspective, the haptic tracks may be mixed to minimize tactile artifacts and ensure a smooth transition.**
- 2. Haptics associated with video media – One or more haptic tracks may be associated with the background video media**
 - a. A global haptic track may be associated with the video media so that it plays continuously during the live event**
 - b. Two or more haptic tracks may be associated with the video media, for example, one haptic track associated with crowd noise, and one associated with on-field noise. A gradual transition may be made between these tracks based on whether a segment of video media is included in the user’s field of view or excluded from the user’s field of view.**
- 3. Haptics associated with a capture device – A haptic track may be associated with a capture device such as a sensor or camera that captures Tactile Essence (SMPTE st2100-1-2017):**
 - a. This haptic track can be activated, deactivated, or modulated based on the user’s viewpoint.**
 - i. For example, if the haptic track is generated by an inertial sensor embedded in a baseball bat, the haptic track may play always, or only if the bat is in the field**

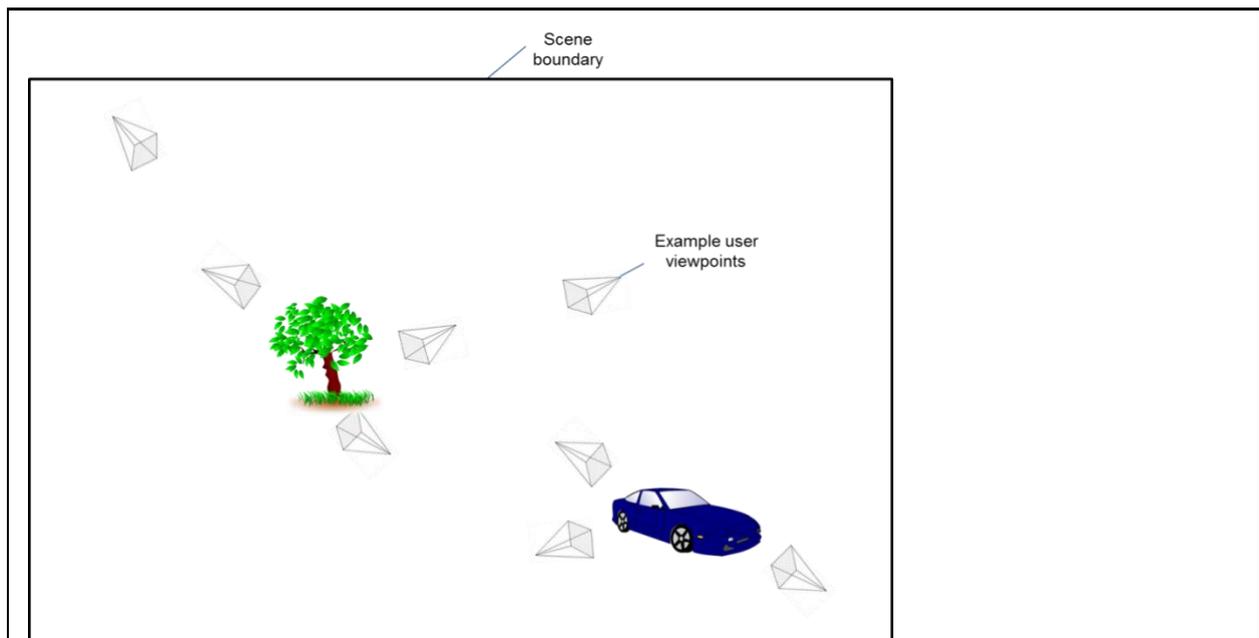


Figure 2: An example of full 6DoF immersive content where a viewer may change his or her viewpoint to any location within the scene boundary

Haptic schema

- The haptic schema for this use case includes most of the elements described in Use case 2.1, with the following additions/modifications:
 - Haptic profiles are associated with objects.

Alternatively, the surface features may be derived from other object attributes such as its surface geometry, applied textures/shaders, or virtual material, visual appearance, context, past interactions, etc.

Potential Haptics Requirements and Specifications

- The specification shall support dynamic generation and synthesis of haptic effects based on other metadata, media streams or external data sources.
- The specification shall support high-fidelity, low-latency user tracking to enable active interaction and haptic feedback

2.3. A Mobile subject looks around a statue Object with up-close views

A single person moves around a still statue object in a room and looks at the statue, with the capability to look closely.

Haptic schema

A haptic profile is associated with the statue. This profile may have different resolutions based on proximity of the user's avatar.

The user may engage with the haptic profiles in different ways depending on the interaction model of the player

Touching part of the screen that contains the visual elements with which the haptic profile is associated
Using a "virtual probe" or "laser pointer", common in VR interaction: an on-screen indicator similar to a cursor that shows where in the scene the user is pointing.

For example, a statue depicted on a sphere may have a gross shape, surface features, and a fine material texture.

- As the viewer interacts with the statue from a distance, haptic effects may be associated with the gross shape of the statue may be used.
- After the viewer gets closer and is able to see visual detail of the statue's surface features, haptic effects associated with higher resolution surface features may be used.

Required features

The media content consists of multiple nested spheres with the same center, and the object is at the center of the spheres, the subject looks at the object from outside of the spheres and has the 3 rotational DoFs plus 2 translational DoFs (no movement in the z-axis). The subject is capable of view changing from sphere to sphere.

Potential Haptics Requirements and Specifications

- The specification shall support association of haptic feedback (tactile, kinesthetic, essence, texture) with a 2D, 3D, AR, VR, audio and/or video objects and environments
- The specification shall support interactivity models related to avatar position and orientation.
- The specification shall support high-fidelity, low-latency user tracking to enable active interaction and haptic feedback

2.4. Entertainment/Immersive Video/Theatrical

This use case refers to the extension of current movie theaters, amusement parks or home theaters to increase the user experience with the addition of physical feedback.

The so-called 4D-cinema are targeting those applications with moving platforms and other devices (fan, heat, water spray...). The 4DX commercial solution from CJ 4DPLEX for cinema, D-Box for cinema and home, TRIOTECH for amusement parks... are some examples of how the technology can be used for entertainment.

More advanced experiences are also proposed to the users with virtual reality (VR) experiences, especially the location-based entertainment (LBE) and family entertainment centers (FECs), as proposed by dimension, the VOID, VRstudios, IMAX, Springboard VR.

The user is sitting on special moving platforms in a theater or at home or wearing special suits in order to feel the created haptic experience. The movie plays and the special haptic effects are played back at the right time during the movie. In LBEs the interaction and immersion are much higher since the user is able to freely move inside the VR scene and interact with it.

Different types of effect are played (wind, motion, water spray, heat, light, impact, touch).

Haptic schema

The haptic effects are captured with dedicated sensors during the shooting of the scenes (for instance the motion using an accelerometer) and ingested for post-processing within an editing tool, or manually synthesized by an artist (haptographer). This is generally done during the post-production stage of the movie (or experience) with appropriate association with the audio-visual content (for instance fade-in/face-out and synchronization).

The user is generally feeling the haptic feedback through dedicated devices such as motion platforms, controllers, suits, etc.

The effects/Haptic data are:

- Motion (acceleration/speed) of the user in case of FPV
- Position and Motion of the camera to move in the scene with a predefined path
- Feedback from the scene such as vibrations (to simulate an earthquake) or wind
- Feedback from another virtual or real character, such as distant interactions or impact from the character to the user
- Feedback from the interaction with an object as depicted in other scenario (impact/force, velocity, physical properties...)

The experience is mostly passive for movie or home theaters, it can be partially or totally interactive for amusement parks, location-based entertainment, and family entertainment centers. Effects are pre-recorded and played back when appropriate. Thus, it addresses Phase 2A, even if some simple sub-cases could be compatible with Phase 1 requirements.

The experience can be much more interactive for amusement parks, location-based entertainment, and family entertainment centers. Effects are created and triggered depending on the user's actions and position. Thus, it addresses phase 2B.

Overlap with other use cases

- Multiple users in VR environment
- Social TV and VR
- Gaming/E-Sports

Required features

- Synchronization with audio and video
- Adaptation to various end user devices
- Appropriate management of motion effects (fade-in/out)
- Distribution and association of different haptic effects/channels

Potential Haptics Requirements and Specifications

- The specification shall support different media types and various haptic feedback paradigms (pre-rendered, synthesized)
- The specification shall support dynamic generation and synthesis of haptic effects based on other metadata, media streams or external data sources.
- The specification shall support mixing and modulation between and within haptic tracks.
- The specification shall support rendering of haptic feedback on multiple devices across multiple body locations
- The specification shall support synchronization of haptic effects with the audio-visual content
- The specification shall support encoding of the mastering or authoring system specification.
- The specification shall support presentation of haptic media on alternate devices or devices which may have different performance characteristics from the mastering system.

3. Phase 2B Use Cases

The use cases in this section involve some level of contact-driven interactivity with the immersive environment. As with Phase 2A use cases, the requirements at the end of each use case are consistent with those specified in [3].

3.1. VR Video Calling

Panoramic video calling : Alex and Bob are at different physical locations, and they are having a video calling, Alex sends 360 video to Bob while Bob send 2D video cause his devices with limited capabilities (cannot take 360 videos)

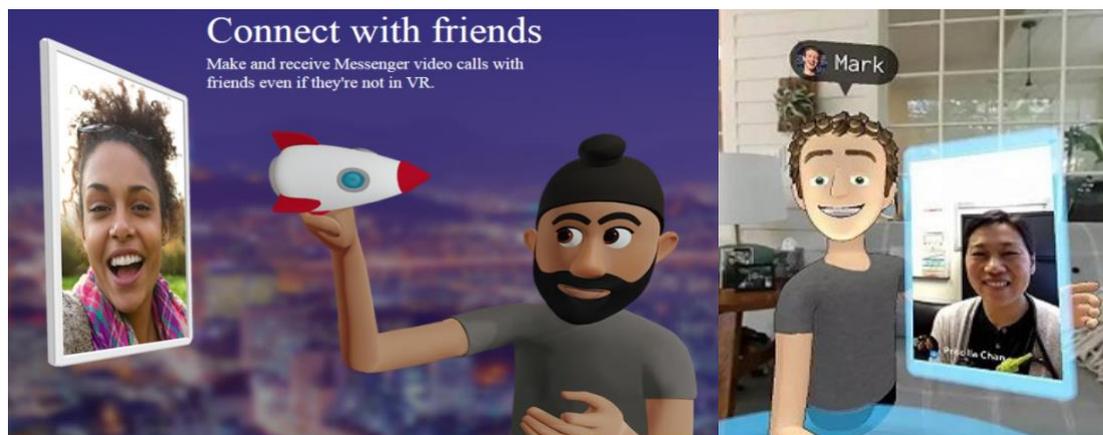
- Bob wears an HMD to watch the 360° video sent by Alex
- If Bob does not have an HMD, he can only play the 360° video on other terminals (mobile phone, PC, etc.)



(<https://www.insta360.com/>)

Alex and Bob in the same VR environment while Bob send 2D video

- Alex can be present in the VR environment through some form of user-embodiment while Bob's 2D video will be present as a VR object (virtual screen, etc.) in the VR environment, which Alex can interact with (move, zoom in/out, etc.)



(<https://www.facebook.com/spaces>)

Haptic schema

Users may send haptic effects to each other in the following ways:

1. Through gesture, for example:
 - a. by touching the user's avatar image (representing the VR user) or the region of the video that includes the image of the person (representing the video user).
 - b. By touching objects in the remote user's environment
2. By attaching and sending an external media element with an associated haptic track or haptic effect such as a haptic sticker, GIF, animation, video clip, or virtual object.

Specialized haptic devices may be needed to experience touch feedback related to manipulation or collision in the 6DOF virtual environment. These devices could include haptic gloves, body suits or specialized controllers.

Overlap with other use cases

- Multiple users in VR environment
- Social TV and VR
- Multiple users in VR environment, 6DoF
- VR Conferencing

Required features

- Interactions with VR objects
- Synchronization of audio and video of users and the scene
- Users whose devices with limited capabilities (without motion tracking) can get into a shared VR environment

Potential Haptics Requirements and Specifications

- The specification shall support association of haptic feedback (tactile, kinesthetic, essence, texture) with a 2D, 3D, AR, VR, audio and/or video objects and environments.
- The specification shall support interactivity models related to avatar position and orientation.
- The specification shall support rendering of haptic feedback on multiple devices across multiple body locations

3.2. Virtual Prototyping

Efficient digital prototyping involves simulating objects' behavior under real world operating conditions; therefore, sensorial experience is as important as the visual one. While VR makes it possible to model systems and interactions visually, haptics may make these interactions immersive. This will increase the perceived realism of the user and provide more valid prototyping insights.

An industrial designer is designing the interior of a new production vehicle using CAD software. Once the interior is mocked up, the designer can put on a VR headset and interact with the interior controls and HMI as well as understand sight lines and reachability. The designer can iterate their design without creating physical prototypes allowing for both more design ideas and lower development costs.

Once the designer has a final design, they can invite other stakeholders into a shared virtual environment that allows the stakeholders to walk around the CAD model, touch it and experience the design with all their senses.

Haptic schema

The user will have a fully tracked virtual avatar in the virtual environment that is able to interact with the virtual media. This likely includes haptic gloves, body suit or other tracking and feedback devices.

Users may experience haptic sensations during interaction with the virtual prototype. The user will be represented by a virtual avatar that has collision geometry suitable for sensing contact with the CAD model and appropriate haptic sensations in time and space will be synthesized by the presentation engine.

- The haptic profile of the virtual objects may be based on the 3D object's geometry. For example, when part of a user's avatar interacts with the object by colliding with it, a haptic effect may be displayed that:
 - Signifies the collision
 - Prevents the user's interaction gesture from crossing the boundary of the 3D object
 - The haptic profile may be based on the 3D object's surface features.
 - These surface features may be explicitly defined at design time such as in the case with a haptic texture being associated with a 3D object

Overlap with other use cases

- Multiple users in VR environment
- Multiple users in VR environment, 6DoF
- VR Conferencing

Required features

- Interactions with VR objects
- Synchronization of audio and video of users and the scene
- Users whose devices with limited capabilities (without motion tracking) can get into a shared VR environment

Potential Haptics Requirements and Specifications

- The specification shall support dynamic generation and synthesis of haptic effects based on other metadata, media streams or external data sources.
- The specification shall support association of haptic feedback (tactile, kinesthetic, essence, texture) with a 2D, 3D, AR, VR, audio and/or video objects and environments
- The specification shall support interactivity models related to avatar position and orientation.
- The specification shall support coding and presentation of interactivity models related to avatar-scene or avatar-avatar interactions.
- The specification shall support rendering of haptic feedback on multiple devices across multiple body locations

3.3. Training and Simulation

VR training (VRT) defines a solution using virtual or mixed reality to transfer useful skills towards the trainee with the use of an extended reality solution. VR training applications typical have the following goals:

- Reduce operational costs of delivering practical training:
 - Reduce real equipment usage for training purposes
 - Reduce maintenance cost for training centers
 - Reduce travel cost for trainee
- Ensure repeatability thanks to the digital format of the learning support
- Ensure learning consistency thanks to the digital support
- Increase user retention through increased immersion and gamification techniques

Typical Use case:

A well-known electrical equipment provider commercializes low and medium voltage electrical equipment needing maintenance on a regular basis. The customer's workforce needs to perform scheduled training to learn and refresh the procedures to perform the maintenance operations. This training is performed at training centers around the globe. The sessions are extremely expensive, involving the travel of the workforce to the training center, a few days stay to perform training activities on dummy machines under the supervision of the trainer. The electrical equipment manufacturer developed a haptics training solution to digitize the maintenance and security training for the workforce to bring the training sessions to the customer. The solution also commercializes a VR training system, including haptics and VR equipment, to allow their customer to keep the training scenarios and experience as documentation.

Haptics enhances the experience as follows:

Skill transfer: including haptic feedback in VR training can generate positive learning reinforcement, enhancing the effectiveness of training.

Realism: The absence of natural interaction and realistic haptics can generate bad practices or negative learning that must be unlearned in real life skill implementation.

Immersion: User immersion increases the embodiment and believability of the training scenario, increasing its effectiveness.

User Experience: Hand tracking and natural interactions coupled with well-designed haptic feedback can meet or exceed user expectations of interactive content.

Haptic schema

The user will have a fully tracked virtual avatar in the virtual environment that is able to interact with the virtual media. This likely includes haptic gloves, body suit or other tracking and feedback devices.

Users may experience haptic sensations during interaction with the virtual prototype. The user will be represented by a virtual avatar that has collision geometry suitable for sensing contact with the VR model and appropriate haptic sensations in time and space will be synthesized by the presentation engine.

Users may experience haptic sensations associated with training goals and other metadata. This feedback may be user and scenario state dependent.

Overlap with other use cases

- Multiple users in VR environment
- Multiple users in VR environment, 6DoF
- VR Conferencing

Required features

- Interactions with VR objects
- Synchronization of audio and video of users and the scene
- Users whose devices with limited capabilities (without motion tracking) can get into a shared VR environment

Potential Haptics Requirements and Specifications

- The specification shall support dynamic generation and synthesis of haptic effects based on other metadata, media streams or external data sources.
- The specification shall support association of haptic feedback (tactile, kinesthetic, essence, texture) with a 2D, 3D, AR, VR, audio and/or video objects and environments
- The specification shall support interactivity models related to avatar position and orientation.
- The specification shall support coding and presentation of interactivity models related to avatar-scene or avatar-avatar interactions.
- The specification shall support rendering of haptic feedback on multiple devices across multiple body locations

3.4. Telerobotic surgery

A user may be a specialist surgeon or surgical resident that is remotely located from a patient in need of urgent surgical treatment. Due to time and distance constraints the surgeon can use a teleoperation system in which they feel a sense of remote presence at the actual surgical site. The surgeon may have a remote robotic assembly able to track their motion and provide tactile feedback during the surgical procedure.

The surgeon needs to have extremely high fidelity visual and tactile feedback about the remote site, which can be provided by a two-way MPEG coded media stream. The downlink media stream should provide extremely high quality visual and tactile data at low latency. The uplink media should provide motion information to the remote robot. Round trip latency should be bounded, and the system should be mechanically stable.

Haptics enhances the experience as follows:

1. **Increased realism:** Haptic feedback from the telerobot to the human provides a greater sense of realism embodied in the robot, a greater sense of dexterity and proprioceptive presence to the human through the robot, and the familiarity of natural movement and task conduct through haptic feedback which acts as confirming sensations for task conduct and completion.

2. **Increased performance:** Using tele-robotics, one person can be present in many geographically dispersed environments through the deployed robotic avatar.

Safety and control: Sparing first-person exposure to dangerous situations or environments, haptic feedback through robotic sensors provides real-time feedback and natural human interface for task completion with high levels of confidence and familiarity with conduct of the telerobotic avatar.

Telesurgery is one of the representative use cases in the IEEE P1918.1 standard (<REF>). It involves a surgeon using cameras, robotic arms, and tactile and kinesthetic sensors to perform surgery on a remote patient across a high-speed network connection.

Other medical use cases include tele-rehabilitation and tele-mentoring of remote patients.

- Sensed information of surgeon’s hand movements such as position, velocity, and orientation transmitted from host to client. Interaction forces and other sensed data at end-effector (surgical tool) transmitted from client to host.
- Resolution and fidelity of haptic feedback is key to guide the surgeon (bone vs. soft tissue) and complement visual information.

Haptic schema

- The remote site tactile essence should be encoded, transported, decoded, and presented to the operator with minimum latency. If necessary, the tactile essence may need to be transcoded due to sensing/actuation asymmetry.
- The local site haptic feedback should have high transparency and stability relative to channel delays.
- System must have low round-trip latency to ensure stability between local and remote site.
- Sensory substitution can be applied when there is a mismatch between the client sensors and the host actuators.
- If feasible, a supervisory control schema could be deployed in which the local operator receives a subset of the tactile essence or abstracted feedback related to the remote site patient state.

Overlap with other use cases

- VR Conferencing

Required features

- Interactions with VR objects
- Synchronization of audio and video of users and the scene
- Users whose devices with limited capabilities (without motion tracking) can get into a shared VR environment

Potential Haptics Requirements and Specifications

- The specification shall support different media types and various haptic feedback paradigms (pre-rendered, synthesized)
- The specification shall support association of haptic feedback (tactile, kinesthetic, essence, texture) with a 2D, 3D, AR, VR, audio and/or video objects and environments
- The specification shall support interactivity models related to avatar position and orientation.
- The specification shall support coding and presentation of interactivity models related to avatar-scene or avatar-avatar interactions.
- The specification shall support specification of round-trip (input to feedback) latency requirement.

3.5. Marketing/Sales

XR technologies have the capability to showcase products and experiences that are impossible or extremely costly to build. It has applications for trade show experiences, in-store experiences for retail, experiential marketing, and to raise user awareness for specific topics.

- **Engage the user.** Well-designed XR marketing experience generates a deep sense of presence, increasing user retention and engagement during the experience.
- **Increase customer conversion rate.** User immersion within virtual reality marketing applications can increase customer conversion. With XR applications, marketers can enhance user immersion and content interactivity. Marketing applications result in a greater lifelike experience and testing capabilities for customers. This is especially relevant for applications in real estate, or complex products like vehicles, tool machinery, or boats.
- **Speed up the sales process.** The sales of complex or spatially large products can be helped by XR applications. The greater benefits happen when a product requires user testing or live presence to evaluate human factors.
- **Reduce marketing costs.** XR marketing applications can be realized once, and deployed on multiple sites, easily transported, shipped, and deployed in front of the customers. An entire catalogue of complex products can fit in a portable headset.

Haptics enhances the experience as follows:

- Realism/Differentiation of active haptic elements - The user will gain a visceral understanding of how the physical product would behave.
- Immersion - The user will feel more present, and this will create greater customer engagement.

Accessibility - Users can interact naturally with the virtual product as if it were real.

Haptic schema

The user may have a fully tracked virtual avatar in the virtual environment that is able to interact with the virtual media. This likely includes haptic gloves, body suit or other tracking and feedback devices.

Users may experience haptic sensations during interaction with the virtual prototype. The user will be represented by a virtual avatar that has collision geometry suitable for sensing contact with the CAD

model and appropriate haptic sensations in time and space will be synthesized by the presentation engine.

For experiential marketing, users may receive tactile feedback in order to constrain their body motion or provide a sense of the desired experience.

Overlap with other use cases

- Virtual Prototyping/Retail

Required features

- Interactions with VR objects
- Synchronization of audio and video of users and the scene
- Users whose devices with limited capabilities (without motion tracking) can get into a shared VR environment

Potential Haptics Requirements and Specifications

- The specification shall support association of haptic feedback (tactile, kinesthetic, essence, texture) with a 2D, 3D, AR, VR, audio and/or video objects and environments
- The specification shall support interactivity models related to avatar position and orientation.
- The specification shall support high-fidelity, low-latency user tracking to enable active interaction and haptic feedback

3.6. *Gaming/E-Sports*

Gaming is moving from a single user, playing with a game console and gamepad, to collaborative experiences through the cloud and new devices. As an example, the latest Sony PlayStation 5 is a good indication of this trend. It supports immersive 3D audio, 4K HDR image resolution, VR, higher frame rates, and a gamepad with several actuators, including both tactile and kinesthetic modalities ([PS5 DualSense controller](#)). Cloud gaming is announced at some point as for the Xbox Series X. More advanced devices are also coming such as the Tesla Suit, or other vest from bHaptics and Actronika, made of several devices and actuators.

In addition, the same type of architecture can be used to develop live events and e-sport applications. Unlike the traditional gaming experience, e-sport mixes gaming, competition, and live feedback for the audience. There is no doubt that new devices providing the professional users with physical feedback during the game, or for the audience having the ability to sense some of the competition actions will be available.

Finally, live sport events (automobile race, football, basketball, hockey...) broadcasted today could be enhanced with some haptic feedback related to the viewed action. The user can feel some of the player's sensations (either from the environment or from the player himself).

Haptic schema

The user might be represented as an avatar in the game, interacting with other players or game characters. The haptic effects are generated at the rendering depending on the user actions and game storyline. A library of available haptic effects can be created to be selected based on the player interactions.

For live events, the haptic effects are either captured or synthesized to provide the user with the same sensations as the one felt by the real players, or just to increase viewers engagement.

For e-sport the same type of haptic feedback can be provided to the viewers, to feel the same feedback as the one felt by their favorite (selected) gamer. Alternatively, the content creator might prefer adding some special effects to keep the viewers engaged in the spectacle.

Haptics data:

- Sensing information of players movements (such as position, velocity, and orientation)
- Information on the scene and the context (wind, force)
- Interaction with the other player or an object (impact/force, velocity, physical properties...)

While the viewers in e-sport events could satisfy the phase 2.a requirements, gamers certainly require phase 2.b technology to freely interact with the content and other players (real or virtual) during the game.

Overlap with other use cases

- Multiple users in VR environment
- Multiple users in VR environment, 6DoF
- Entertainment
- Broadcast

Required features

- Synchronization with audio and video, and other characters from the action (real or virtual)
- Very low latency to address game interactions
- Adaptation to various end user devices (targeting high-end gaming platforms, live TV shows, mobile)
- Distribution and association of different haptic effects/channels when addressing high-end platforms

Potential Haptics Requirements and Specifications

- The specification shall support different media types and various haptic feedback paradigms (pre-rendered, synthesized)
- The specification shall support dynamic generation and synthesis of haptic effects based on other metadata, media streams or external data sources.
- The specification shall support interactivity models related to avatar position and orientation.
- The specification shall support coding and presentation of interactivity models related to avatar-scene or avatar-avatar interactions.
- The specification shall support rendering of haptic feedback on multiple devices across multiple body locations
- The specification shall support synchronization of haptic effects with the audio-visual content
- The specification shall support high-fidelity, low-latency user tracking to enable active interaction and haptic feedback
- The specification shall support presentation of haptic media on alternate devices or devices which may have different performance characteristics from the mastering system.

3.7. Telepresence

Telepresence is the ability to be virtually in a distant place with other people (real or virtual), interacting and moving with them. The participants are represented through an avatar potentially representing the user with high fidelity or through a virtual synthetic representation. In some advanced cases, the avatar could be physical through the control of a distant robot.

This is exactly the challenge that the ANA Avatar XPRIZE is trying to address ([ANA Avatar XPRIZE | XPRIZE Foundation](#)). Haptic feedback is a central part of the solution to provide a physical telepresence.

This was also particularly identified during the COVID pandemic, where people interactions and socialization have been difficult because of no satisfactory technologies. Some early platforms (such as NextGen Event, vFAIRS, Vmeets, VirtuLab, Hopin, MeetingPlay...), try to provide better solutions, but are still lacking the sense of touch. Technologies such as VRgluv or HaptX provide solutions to touch, grip and feel objects or people in a distant environment.

It applies to communication, social network, entertainment, education, work, conference, and tourism.

Telemedicine or telesurgery is not part of this use case. A special section is dedicated to this one due to the special nature of medical applications (security, very low latency, privacy...).

Haptic schema

The user is represented by his avatar, either realistic or totally imaginary. The user is also tracked to reproduce motion, expressions, or any other gesture. The application is necessary bi-directional to provide feedback from the user action or from another participant. Distant devices represent the user

body partly (i.e., a glove) or totally (humanoid robot), through which interaction with people or objects are possible.

Emotions/empathy are very important for this social relationship, thus realism (or plausibility) of the telepresence is an important factor of adoption.

Haptics data:

- Sensing information of others to socialize (touch, empathy)
- Sensing information on the scene and the context (wind, force)
- Interaction with other people or objects (impact/force, velocity, physical properties, emotions...)

The level of interaction and emotional involvement of this use case require the most advanced technologies from phase 2b.

Overlap with other use cases

- Multiple users in VR environment
- Multiple users in VR environment, 6DoF
- Tele-surgery
- Training and Simulation
- Marketing/sales

Required features

- Synchronization with audio and video, and other characters from the action (real or virtual)
- Very low latency to address control of distant devices
- Adaptation to various end user devices (targeting high-end platforms or mobile)
- Distribution and association of different haptic effects/channels when addressing high-end platforms

Potential Haptics Requirements and Specifications

- The specification shall support different media types and various haptic feedback paradigms (pre-rendered, synthesized)
- The specification shall support dynamic generation and synthesis of haptic effects based on other metadata, media streams or external data sources.
- The specification shall support interactivity models related to avatar position and orientation.
- The specification shall support coding and presentation of interactivity models related to avatar-scene or avatar-avatar interactions.
- The specification shall support rendering of haptic feedback on multiple devices across multiple body locations

- The specification shall support synchronization of haptic effects with the audio-visual content

3.8. Driver Augmentation

Description:

Modern vehicles have numerous sensors and data sources that enable the construction of a spatial map of the surroundings of the vehicle. These spatial maps are utilized for driver assistance and autonomous driving applications. Often the maps consist of Lidar and optical sensor data and may be efficiently represented as point clouds using MPEG VPCC and related technologies. Semi-autonomous driver scenarios commonly provide some type of haptic feedback to the driver as a result of calculations on this point cloud – including steering feedback, driver seat vibration feedback and pedal feedback. In this use case, haptic feedback is intended to provide guidance and assistance to the operator and as such, must have a high level of robustness and physical validity. These features distinguish this use case from other 3D interaction use cases.

Typical use cases:

- **Alerts:** High priority notifications that aim to shift the driver’s attention away from their primary focus and toward the alert in order to provide critical, time-sensitive information. Typical use cases for alerts include collision warnings, time sensitive navigation messages, and cues from driver assistance systems.

Haptic schema

- Another design consideration is body locus. For example, haptic effects embedded in the steering wheel, or the seat can map the directionality of the alert to the side of the driver’s body that corresponds to the event that generated the alert. A haptic effect on the driver’s left side (palm, leg, or back) can thus alert the driver to an event that is occurring on the left side of the car or road. Less critical alerts can also use this schema, for example, navigation cues.

Overlap with other use cases

- Somewhat similar to tele-surgery in the sense that both use cases need a high level of reliability and validity for the haptic interactions.

Required features

- Support for coding of multimodal sensor data including point clouds, range data, and other environmental sensing
- Support for coding of video for machines

Potential Haptics Requirements and Specifications

- The specification shall support rendering of haptic feedback on multiple devices across multiple body locations
- The specification shall support specification of round-trip (input to feedback) latency requirement.
- The specification shall support encoding of the mastering or authoring system specification.

- The specification shall support presentation of haptic media on alternate devices or devices which may have different performance characteristics from the mastering system.

3.9. ATSC 3.0 Broadcasting

One of the promises of ATSC 3.0 is the ability to view broadcast content on mobile devices (A/300 [4]) when they are used either as primary devices (i.e., have an ATSC 3.0 tuner built-in) or as companion devices (A/338 [5]). At the time of writing, most mobile devices in the market have just one haptic actuator, but new devices with multiple actuators are beginning to show up. The key idea: if these haptic actuators can be utilized in synchronization with the audio-visual content, then the viewing experience of ATSC 3.0 content on these mobile devices would be significantly enhanced and new use cases enabled. Some common examples of the types of content that could benefit from the haptic signals:

- Live sports (major league and professional sports) – viewers can feel the game events (such as hits, catches, etc.), in addition to watching and listening to them, enhancing their overall immersive experience of the game.
- Action films/shows – a haptic track can increase the immediacy and impact of explosions, car chases, etc., drawing the viewer into the action.
- Advertisements – studies have shown that haptics-enabled mobile ads improve brand favorability and are more effective in driving purchase intent among viewers.

Details of accomplishing this use case are described in the ATSC Recommended Practice on Haptics for ATSC 3.0 (A/380 [6]).

Haptic Schema

- Depending on the size of the haptic track (that varies by content), the haptic track can either be included in the ATSC 3.0 event stream itself or it can be retrieved from a separate URL along with an authentication token.
- For live sports events, the output from multiple sensors on the field of play are part of the ‘Sensor Feed’ that is uploaded to the Cloud Repository and used to generate the haptic events that become part of the haptic content JSON.

Overlap with Other Use Cases

- Entertainment/Immersive Video/Theatrical

Required Features

- A restful web interface for accessing the haptics file from the cloud repository

- Ability to generate haptic effects from sensor feeds

Potential Haptics Requirements and Specifications

- The specification shall support high-definition haptics
- The specification shall support dynamic generation and synthesis of haptic effects based on other metadata, media streams or external data sources.
- The specification shall support specification of round-trip (input to feedback) latency requirement.

3.10. Apparel Fitting Rooms/Retail

Online shopping has gained in popularity and changed the customer’s experiences. However due to the lack of sensory experience, shoppers may be reluctant to purchase a garment online. It is well known that in traditional retail shopping there is a close relationship between touch, information acquisition, desire development and the motivation to purchase products. Haptic feedback could solve this problem by allowing shoppers to feel a piece of fabric.

As depicted below, most online shopping systems allow to magnify the picture of the product in order to better see the micro details. A haptic system in this context would allow to touch this magnified area.



Image source: Susana C. Silva, Thelma Valeria Rocha, Roberta De Cicco, Renata Fernandes Galhanone, Luiza Tari Manzini Ferreira Mattos, Need for touch and haptic imagery: An investigation in online fashion shopping, Journal of Retailing and Consumer Services, Volume 59, 2021, 102378, ISSN 0969-6989, <https://doi.org/10.1016/j.jretconser.2020.102378>.

The users may then experience haptic sensations during the selection of clothes, and could feel the fabric, its elasticity, and texture. Haptic rendering would be limited to texture and contact. Current hardware cannot render the feeling of grasping a piece of fabric.

Haptic schema

<ol style="list-style-type: none"> 1. Haptics associated with image <ol style="list-style-type: none"> a. haptic data localized in the image. Different zones are possible (see picture: t shirt texture is different from sweater texture) b. Heterogeneous information is available: friction, micro geometry, stiffness, temperature, etc. Multiple tracks are possible c. Scalability: the user can zoom in and out
<p>Overlap with other use cases</p>
<ul style="list-style-type: none"> • Similar to virtual prototyping but focus on texture.
<p>Required features</p>
<ul style="list-style-type: none"> • Haptic interaction with image • The user’s hand is tracked with low latency
<p>Potential Haptics Requirements and Specifications</p>
<ul style="list-style-type: none"> • The specification shall support mixing and modulation between and within haptic tracks. • The specification shall support synchronization of haptic effects with the audio-visual content • The specification shall support high-fidelity, low-latency user tracking to enable active interaction and haptic feedback

References

- [1] ISO/IEC SC29 WG11 (MPEG), *N19513 MPEG-I Phase 2 Haptics Use Cases*, MPEG131, July 2020.
- [2] WG02 20717 N00115, “Haptics Phase 2 – Motivation, Issues, and Use Cases”, MPEG135, July 2021.
- [3] WG02 20947 N00130, “Requirements for MPEG-I Phase 2”, MPEG136, October 2021.
- [4] ATSC: “ATSC Standard: ATSC 3.0 System,” Doc. A/300:2020, Advanced Television Systems Committee, Washington, DC, 15 May 2020.
- [5] ATSC: “ATSC Standard: Companion Device, with Amendment No. 1 (A/338),” Doc. A/338:2019, Advanced Television Systems Committee, Washington, DC, 21 January 2020.
- [6] ATSC: “ATSC Recommended Practice on Haptics for ATSC 3.0 (A/380)”, Doc. A/380:2021, Advanced Television Systems Committee, Washington, DC, 3 February 2021.

Annex A: Haptics in Immersive Media – A Tutorial

Introduction

This annex is intended to provide a foundation to better understand the haptic use cases described in this document.

Haptics provide value to end users in immersive media by engaging the sense of touch. Immersive media endeavors to change a user's sense of place by presenting a convincing sensory illusion. The goal then of such an interface is to occupy and control as much of a person's sensory bandwidth as possible by engaging many afferent nerve endings with high resolution signals. Immersive media that engages only the eyes and the ears can never be as immersive as that which also engages touch, because the sense of touch is a salient element of people's sense of place.

The role of haptics in VR environments has been studied extensively since VR first emerged as an area of technology R&D. Haptics has been definitively shown to enhance end-user assessments of immersion, presence, realism, performance, emotional state, among others.

Value of Immersive Content with Haptics

By providing access to an additional sensory channel, haptic technology provides content creators with a powerful tool to create differentiated and highly engaging user experiences. The variety and depth of content are increased when haptics is available in the content creator's design palette.

When choosing which elements of an immersive experience to include in a piece of content, creators must account for the availability of devices capable of rendering immersive content. Fortunately, almost all devices targeted for distribution of immersive content have some amount of haptic capability built in.

Mobile handsets have haptic actuators that are increasing in quality year over year, now allowing for rich, dynamic vibrations that are closer in experience to a texture than to a buzz. The APIs and SDKs that enable haptic apps and content are increasing in number and sophistication.

Console game and VR controllers are also adopting advanced haptic features. Dual-motor rumble feedback, the standard for many years, has given way to HD vibration actuators and force impulse actuators, engaging more nerve endings in the skin and muscles of the hand. Interaction design patterns for VR user interfaces have come to rely on haptic confirmation of user actions, and developers are increasingly incorporating advance haptics into games to enhance immersion.

Typology of Haptic Immersive Content

Haptic immersive media may be divided into four types that relate the content type to a preferred haptic design approach.

Story-driven Content

In story-driven content, the immersive experience draws the viewer along a narrative storyline similar to linear 2D content like TV shows and movies. In this type of content, the camera is likely to move. While the viewer may have the freedom to view the scene in any direction in

360 degrees, there is one focal point of action that is assumed to be the center of attention. This focal area is where the action of the story takes place – for example, a child playing in a park, and other actions consist of a background that helps create a sense of immersion – for example, leaves in the surrounding trees moving with the wind.

An example of this type of content is HELP¹, an immersive film by 360 Google Spotlight Stories that follows a classic dramatic arc.

In this type of content, haptics will follow the action. Salient events like explosions and crashes will have haptic effects that enhance the drama and intensity of the experience. Haptics may also be used to accentuate moments when a scene turns, or a significant storyline development takes place. The actions of main characters will be prioritized. The design approach will convey “cinematic realism,” where actions and their haptic correlates may be exaggerated for dramatic effect. Haptic effects may correlate to action that is off axis from the expected viewing angle, but this too is used as a story element that prompts the viewer to look in another direction to discover some new story element.

Experience-driven Content

In experience-driven content, the design goal is to make the viewer feel like they are inside the action of the scene. A story is usually not required, because the end-user value of the content is about feeling present in an exciting or unusual reality. All viewing angles are important, because part of the appeal of the content is that the user can suspend their disbelief by looking in any direction and experience a complete illusion.

An example of this type of content is Experience the Blue Angels² by USA Today, which makes you feel as if you are inside an aircraft in an aerobatic jet squadron.

In this type of content, haptic effects will be used to heighten the viewer’s sense of immersion without prioritizing one viewing angle over another. Haptic actions will affect the entire virtual body of the perspective taker, for example, the vibration and shake of a vehicle being ridden by the viewer.

Event-driven Content

In event-driven content, the design goal is to make the viewer feel present at a single location where the action is unfolding.

In this type of content, focus on a specific part of the scene is so critical that it is not always necessary to provide detailed environmental cues. In fact, some of this content limits viewing angles to 180 degrees. The action often comprises a sporting event or artistic performance that takes place on a playing field or stage.

An example of this type of content is NextVR³, a cross-platform app that provides live event video feeds for VR headsets.

¹ <https://www.youtube.com/watch?v=G-XZhKqQAHU>

² <https://www.youtube.com/watch?v=H6SsB3JYqQg>

³ https://www.youtube.com/watch?v=CY_wiN626ac

In this content type, haptics heightens the excitement of the event by providing ambience. The roar of a crowd and the shake of the stands would be candidates for haptic effects.

Interaction-driven Content

In interactive content, the viewer is a participant in the scene. A controller, gestural interface, or other mechanism of engaging with the content is provided to the user. Scenes are often rendered with 3D game engines, and a key user desire is to be able to freely interact with the environment and characters.

In this type of content, haptic effects are usually designed to simulate interaction of content elements with the user's body. Even if an event is very salient, if such an event would not give rise to touch sensations in real life, they would not have haptic effects associated with them in the content.

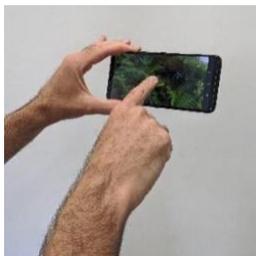
An exception is made for haptics related to usability. For example, a common design pattern in VR is for the user to use a light beam pointer to select menu items, with haptic pulses confirming when a target is acquired, or a button is activated. In this case, haptic effects are not related to immersion in the scene, but nonetheless play a key role in supporting the user's comfort and performance during the experience.

Content Consumption Modalities

Immersive media can be consumed in several ways. Creating haptic tracks for spherical media must account for multiple modalities of interaction with spatial video and interactive content. There are three primary interaction modalities, and in each one, different parts of the user's body are in contact with the playback device, so each modality implies its own set of haptic design criteria.

The challenge with this situation is that immersive content player software can operate in any of the three modalities. Therefore, it may be often necessary for the immersive content to contain multiple haptic tracks, each one created for a different consumption modality, and for the player to select between the tracks based on currently active content consumption modality.

Finger Gesture



The first, most common interaction modality is finger gesture interaction. The orientation of the viewer's point of view is controlled by dragging the finger to reorient the view. The finger may also be used to touch, tap, or otherwise interact with objects in the scene.

In this modality, haptic effects are felt in both the supporting hand and the fingertip of the primary hand. Because more surface area of the supporting hand is in contact with the device, more nerve endings are stimulated, and the sensation in the supporting hand dominates the haptic experience. At the same time, the conscious attention paid to the location of the fingertip, in terms of both finger gesture and visual focus, often cause haptic effects to be correlated to finger actions. The user's correlation of haptics to the background scene and the foreground interaction is unstable and context-dependent, necessitating a design approach that considers the use case.

Hand Gesture



Immersive content may also be experienced by orienting the entire device to control the viewing angle. The orientation of the user's point of view is controlled by moving the device to "scan" the environment.

In this modality, both hands are generally engaged by haptic effects, which makes haptics readily interpretable when they relate to the environment or ambience of the scene. For example, in lived reality, if a person is near an explosion, there would be a resultant vibration that might be felt in nerve endings throughout the body, regardless of whether the person was looking directly at the explosion when it occurred. Thus, to make the experience of an explosion more engaging in immersive content, a haptic effect can be played that synchronizes with explosion event regardless of the user's viewing angle at the moment the explosion occurs. In fact, some content uses this as part of the story by using offscreen haptics to prompt the user to reorient their device and find the source of the haptic event.

In other cases, for example when a scene has a lot of action, creating haptic effects for all the events in the sphere would confuse the viewer and overwhelm their sense of touch. In these cases, it is better to design haptics that correlate to only certain parts of the action, for example, only the most prominent action that is included within the current viewing angle.

Head Gesture



That same immersive content may additionally be experienced with a headset. In this case, the orientation of the body and head are used to shift viewing angles. In some cases, the headset itself has integrated haptic actuators, allowing the person to feel touch sensations on the face, through the eye mask, or head, through the strap. The user may or may not also be holding a controller, or a pair of controllers.

In this modality, haptics played at the headset might be used to reinforce presence in an environment, for example, weather patterns like rain or snow. If near an explosion, haptics at the headset might represent both the air pressure gradient and a light peppering of dirt particles. In both these cases, the haptic effects are attempting to simulate what the person's head would be feeling if they were really in the scene.

At the same time, haptic effects related to manual gestures like pointing, selecting, or interacting with virtual objects could be played in the handheld controllers.

Dimensions of Haptic Design

Haptic designers work with a set of design dimensions that map to the overlap between haptic playback technology capability and human interpretation of touch stimuli. Audio designers work with concepts like volume, panning, and EQ, and have tools that provide access to these design dimensions. Haptic designers have similar practices.

Here are some key haptic dimensions for interactive content:

- Intensity, spanning from light to strong.
- Crossfade curves.
- Timbre, which includes frequency and waveform for high bandwidth actuators.



**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: 2022-04-29

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

Expected action: None

Action due date: None

No. of pages: 17 (with cover page)

Email of Convenor: igor.curcio@nokia.com

Committee URL: <https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-2>

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 2
MPEG TECHNICAL REQUIREMENTS**

ISO/IEC JTC 1/SC 29/WG 2 N00190

Online - April 2022

Title	Use cases and Requirements for Video Coding for Machines
Source	WG 2, MPEG Technical Requirements
Status	Approved
Serial Number	21545

Table of Contents

1	Terms and Definitions	3
2	Introduction	4
2.1	Motivation	4
2.2	Scope	5
2.3	System Overview	5
3	Use cases	7
3.1	Surveillance	7
3.2	Intelligent Transportation	7
3.3	Smart City	9
3.4	Intelligent Industry	9
3.5	Intelligent Content	10
3.6	Consumer Electronics	12
3.7	VCM multi-task with descriptors	12
4	Requirements	14
Annex A Additional Use cases		16
A.1.	Machine vision use case list	16
A.2.	Hybrid human and machine vision use case list	17

1 Terms and Definitions

1. **VCM**: a process of encoding and decoding video, descriptor or feature extracted from video for a machine task.
2. **Machine task type**: a type of computer task that can be applied to image/video or associated data, such as image classification, object detection, instance segmentation, or object tracking.
3. **Machine task**: a specific instance of a machine task type, e.g., an object detection optimized for person detection
4. **Feature**: data representation for a machine task including intermediate output of a network layer
5. **Image classification**: determining whether an object of interest is present within an image or video frame (e.g., deriving a class label for an image)
6. **Object detection**: determining the locations of one or more objects of interest within an image or video frame (e.g., rectangular bounding box with label of the class the object belongs to)

7. **Instance segmentation:** identifying the boundaries and the regions occupied by one or more objects within an image or video frame (including label of the class the object belongs to)
8. **Object tracking:** determining the motion trajectory of one or more objects by locating the position of each object in subsequent video frames
9. **Pose estimation:** identification of posture and/or appearance of an object or person by locating key points or key features (e.g., joints, silhouette, shape, orientation) and/or tracking movement (e.g., gait) and using these to assign a label to pose.
10. **Action recognition:** determining if an action is being performed by a person and identifying what action is being performed by assigning a label to the detected action.
11. **Multi-task:** more than one machine task that are performed in series or in parallel on the output of a VCM decoder
12. **Machine analysis:** a process performing a machine task utilizing the output of the VCM decoder
13. **Neural Network task:** a process performing a machine task implemented using a neural network, which may be divided into feature extraction (part 1) and the remainder of the neural network (part 2)
14. **Human consumption:** a process of using the output of the VCM decoder directly by humans
15. **Feature extraction:** feature extraction is a common sub-process in image or video analysis by which signal values are subject to transformations that yield features of the signal.
16. **Feature encoding:** a process of transforming features into a binary representation of any form
17. **Reconstructed Data:** decompressed output of the VCM decoder to be consumed by a machine or a human (e.g., video, features, descriptors or combinations thereof)
18. **BD-rate:** a method to compare the performance of two curves of bitrate and machine task performance

2 Introduction

2.1 Motivation

Traditional coding methods aim for the best video under certain bit-rate constraints 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.

2.2 Scope

MPEG-VCM aims to define a bitstream from encoding video, descriptors or features extracted from video that is efficient in terms of bitrate and performance of a machine task after decoding.

2.3 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 the features should be specified. Features may take different forms.

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 feature or 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.

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 are not required to be based on deep learning;

VCM can achieve analysis efficiency, computational offloading and privacy protection as well as compression efficiency, while traditional video coding pursues mainly on compression efficiency.

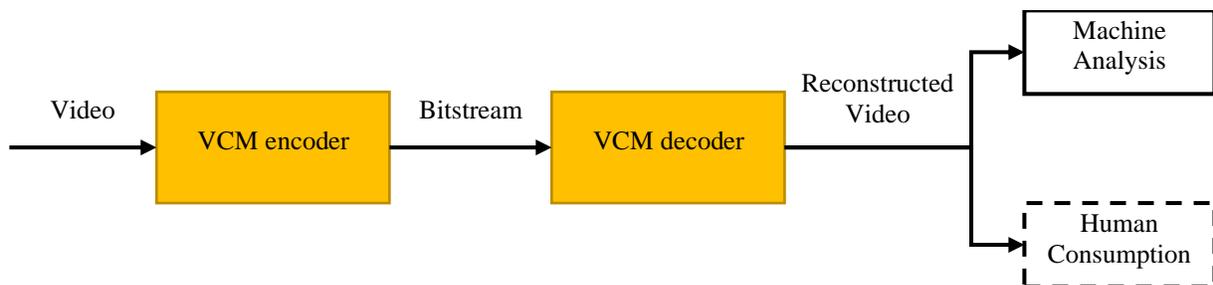


Figure 1-a

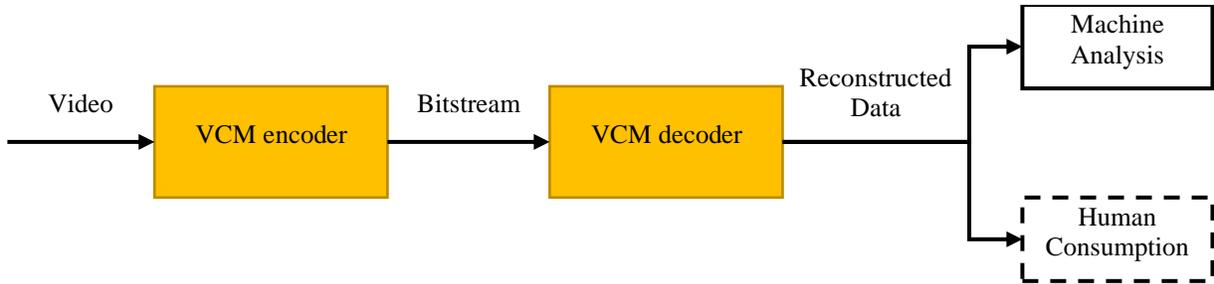


Figure 1-b

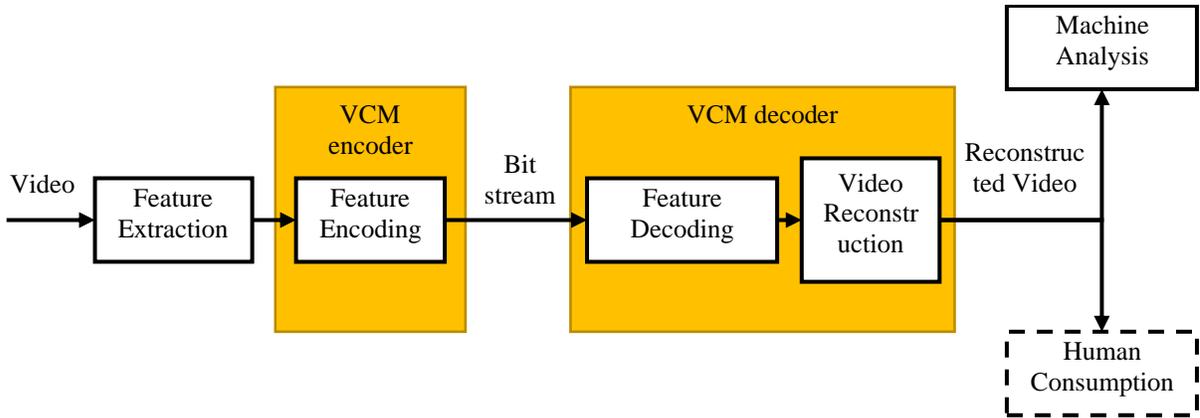


Figure 1-c

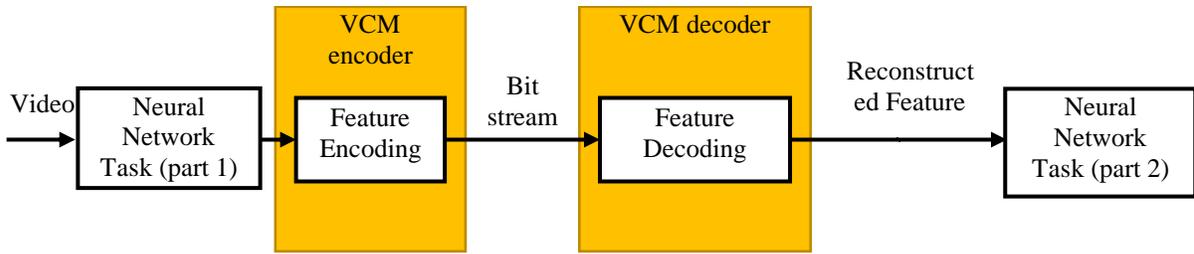


Figure 1-d

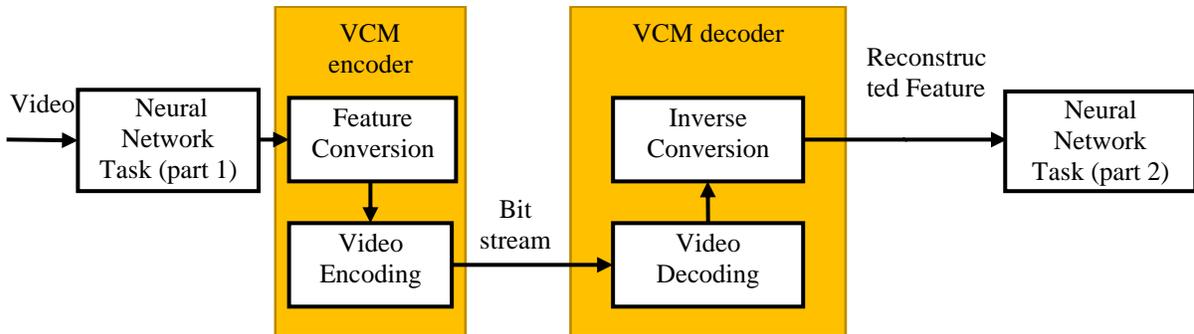


Figure 1-e (Note: video encoding and decoding with existing standard)

Figure 1. Examples of possible VCM architectures

Figure 1 shows five examples of possible VCM architectures. The VCM decoder could include a video decoder, a feature decoder, a descriptor decoder, or combinations thereof.

3 Use cases

In the following subsections different use cases where a VCM standard may be applied are described.

3.1 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 bandwidth for transmission due to the number of sensors and length of video to be transmitted. Besides, the increase in the number of front-end cameras adds heavy computation loads to the back-end server where intelligent tasks are performed.

3.1.1 Intelligent Kitchen

Intelligent kitchen is one of the sub-use cases of Surveillance. It helps detect events that are potentially dangerous in the kitchen and helps ensure a safer and cleaner kitchen environment. Multiple tasks could be performed in this scenario. Action recognition is performed to prevent smoking in the kitchen, outfit recognition is performed to make sure all staff members are wearing their hats and masks properly, motion detection is performed to prevent intrusion to the kitchen after the restaurant is closed. Other than the intelligent tasks mentioned above, many more tasks could be implemented like rat detection, fire detection, gas tank detection, and beyond.

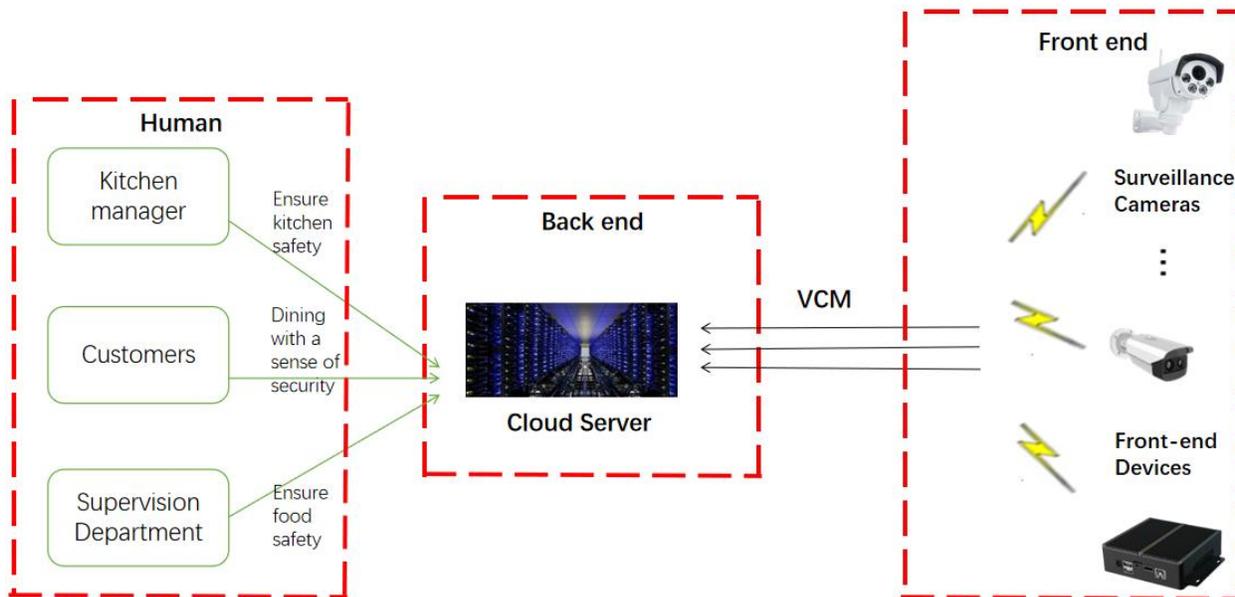


Figure 2. Pipeline of intelligent kitchen use case

Some tasks may share common up-stream tasks or pre-processing like object detection or descriptor extraction. Then, descriptors could be extracted on the front-end devices and sent to

back-end server along with videos/images/features encoded by VCM codec. This could be achieved by the pipeline in Figure 1-b.

3.2 Intelligent Transportation

In a 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.

3.2.1 Cooperative and Connected Vehicles

Cooperative and connected vehicles stand at the center of the emerging VCM standard in intelligent transportation systems. Internet of Vehicles (IoV) is expected to play a key role in the future of urban transportation systems, as they offer the potential for additional safety, increased productivity, greater accessibility, and better road efficiency.

As an example, consider a use case scenario shown in Figure 3. The front car with multiple cameras can see the surrounding environment and detect and recognize objects such as cars, pedestrians, or street furniture or even recognize events such as traffic jams or accidents using (deep) neural networks. The processed data (feature maps) can be consumed internally for desired tasks and/or the extracted features can be compressed and transmitted through the 5G-V2X standard to other surrounding cars/infrastructure (e.g., side road cell/grid) for further analysis. Sending a standardized compact bitstream is essential for interoperability between various vendors, IoV, and IoT applications.

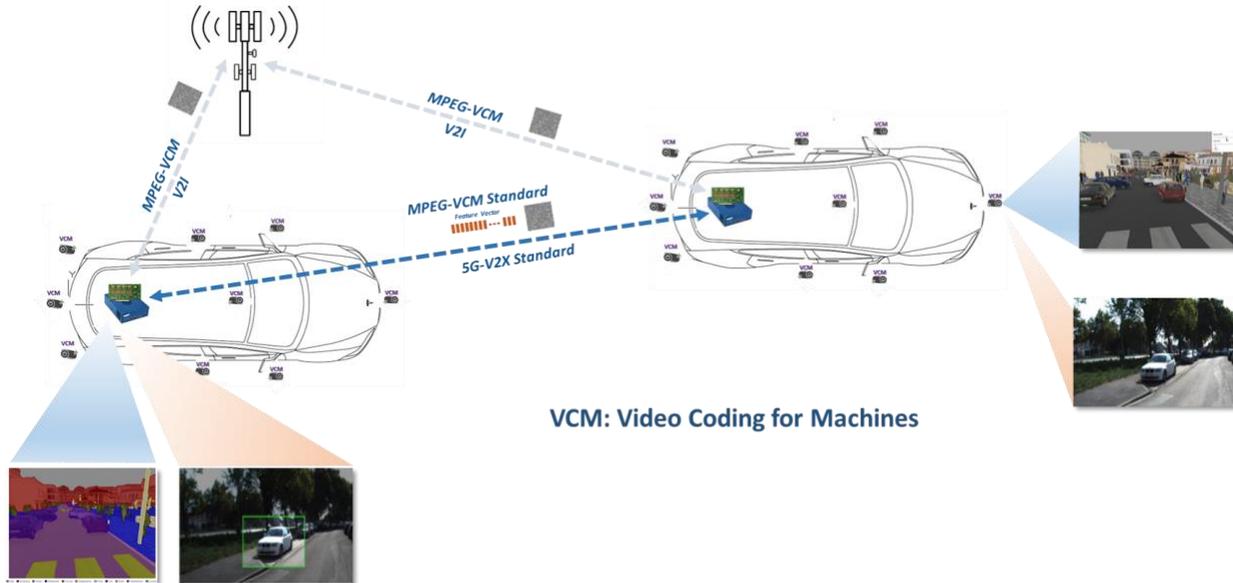


Figure 3. Cooperative V2X-VCM

3.2.2 Feature from additional input channel

Another scenario for consideration is the transmission of features extracted from additional input channels such as invisible light. Detection of invisible or hidden objects during nighttime, fog, or rainy conditions is one of the most 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 various difficult conditions. Furthermore, emerging ICT technologies enable communications among vehicles and infrastructures in the mobile environment. 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 using the data for machine-oriented tasks such as path planning and/or human-oriented tasks overlaid detection results on the video for assisted driving.

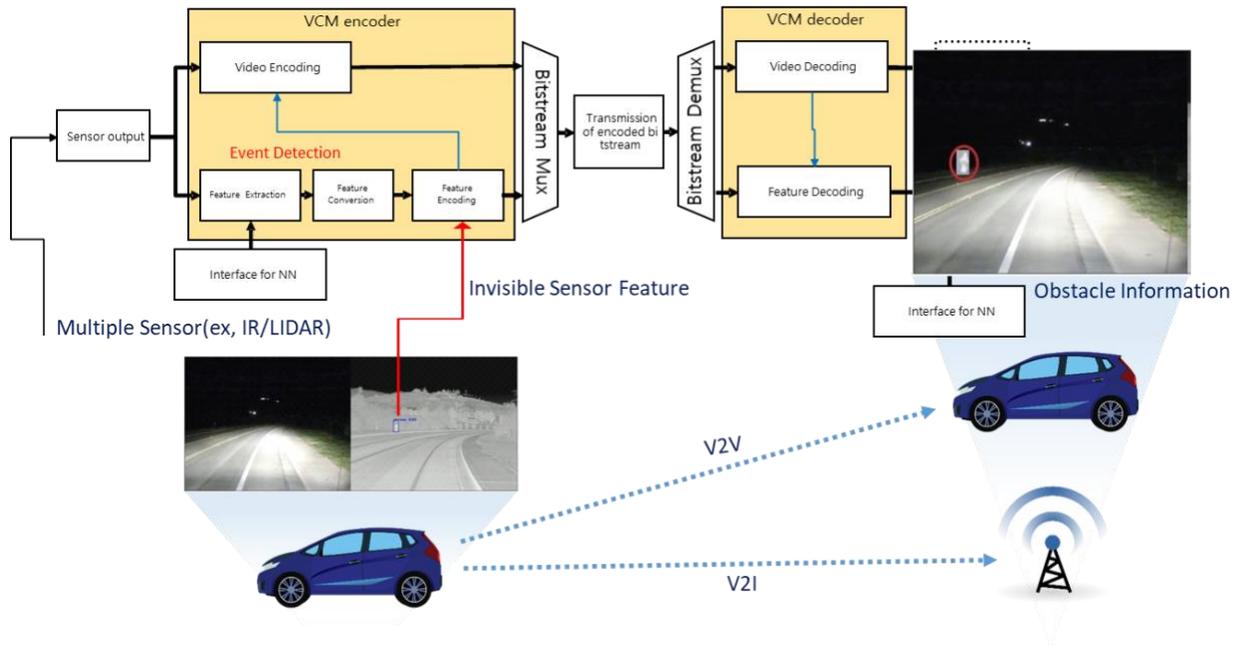


Figure 4. Hidden/invisible object alarm

3.3 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.

3.4 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.

3.4.1 Steel Plate Defects Detection

Steel plate defects detection is a typical sub-use case of intelligent industry. In this scenario, high precision industry cameras take pictures of the steel plates while they slowly moving on the assembly line until all plates are covered. Then, Pictures are pre-processed and encoded before sent to the back end for decoding and detection. At this stage, features are extracted by a backbone network, and encoded by the VTM feature encoder (following Figure 1-d). After the bitstream is received at the backend, they are decoded and fed into task network for analysis. The frequency that the pictures are taken is low because it is usually capped by the time complexity of the traditional codec. And this condition could be potentially improved with the help of low complexity VCM codec.

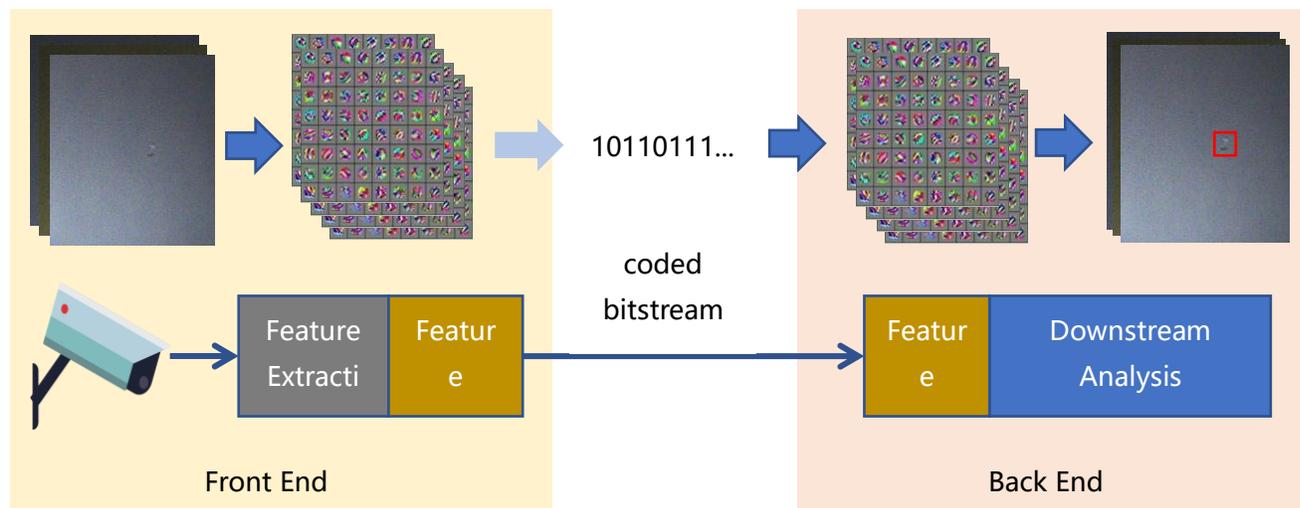


Figure 5. Pipeline of steel plate defect detection

3.5 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.

3.5.1 Video Concealment

These days CCTVs are widely used to prevent crimes or provide useful information such as regional or traffic information. However, personal privacy infringement has been a big issue in the security field. 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, a VCM encoder detects a private area in each scene, and then encodes and transmits the original video after masking the private

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.

- Privacy Video Processing

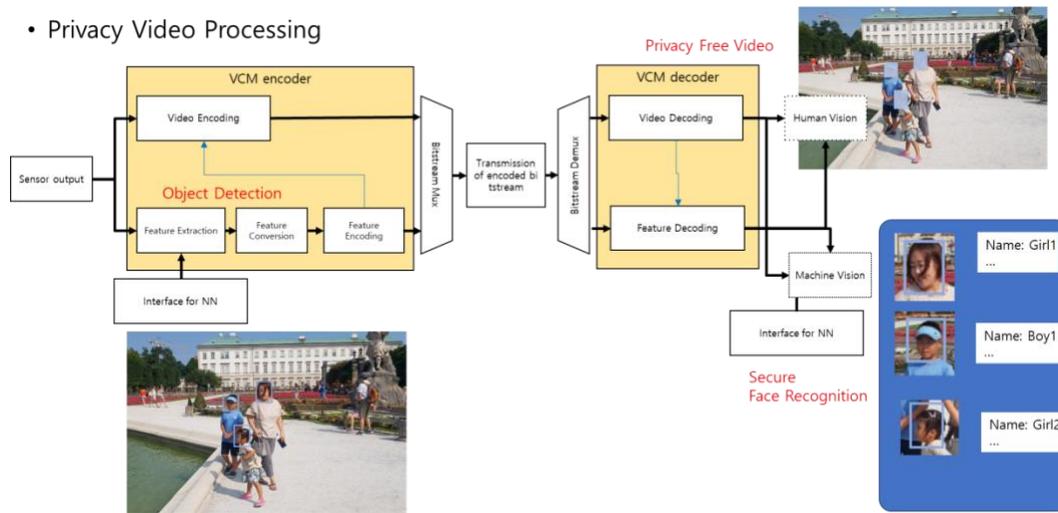


Figure 6. Privacy control in video surveillance

3.5.2 Content Detection

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

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

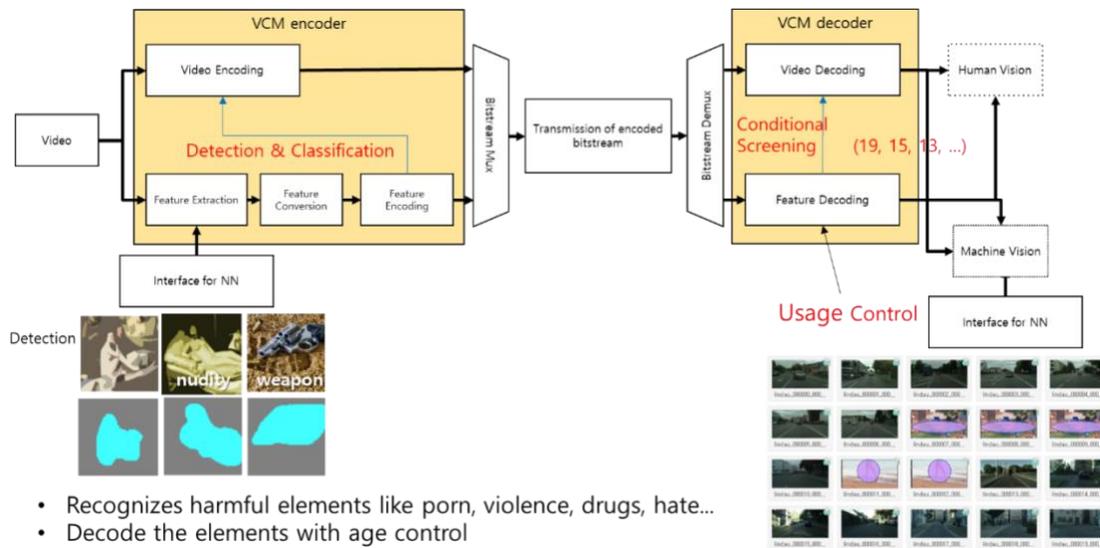


Figure 7. Content protection

3.6 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.

3.7 VCM multi-task with descriptors

A single video stream captured by a surveillance camera can be used for a various kind of machine tasks for different purposes; identify a person or an object, track a person or an object, identify treats, etc. Figure 8 shows a multi-task architecture with descriptor example drawn from the pipeline example shown in Figure 1-b. As shown in Figure 8, each of the different machine tasks can be run in parallel, in series, or in mixed fashion. However, in order to save bandwidth to transmit the captured video and computation time of machine tasks after decoding, a machine task of which the output is commonly used as input to the other machine tasks can be performed before transmission.

For example, if a surveillance application aims to identify and track a person or identify an object being in possession of a person, the common input or preliminary input for such machine tasks is 'person'. Therefore, the system needs to detect 'person' using a machine task and transmit only portions of the detected 'person' in the original video to save bandwidth.

In case of simple 'person' tracking, since an object detection task of 'person' may produce descriptors of detected person e.g., bounding box information, confidence level, id number, etc., the system only needs to utilize transmitted descriptors to track a person and does not need to re-detect 'person' from the decoded video, and hence, computation time and resources are reduced.

Furthermore, if the additional machine tasks require decoded video as input, the descriptor can reduce computational time and resources. For example, if an additional machine task type is a gait detection to identify a person with a walking disability, the system can extract the ‘person’ portion from the decoded video using a descriptor instead of re-detecting ‘person’. The following sub-clauses gives step-by-step examples of a use case on VCM multi-task use case with descriptors.

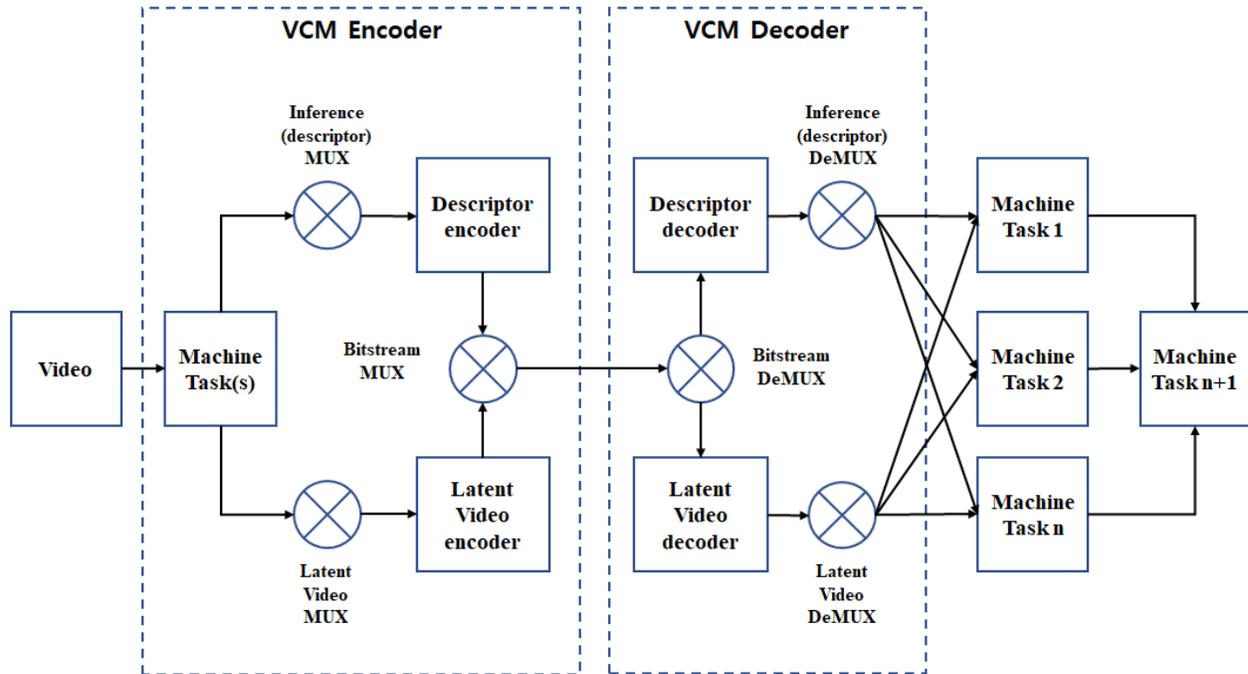


Figure 8. Example of VCM multi-tak with descriptor

3.7.1 Preliminary operation (common operation)

- a. The VCM encoder operation
 - A surveillance video is fed into the VCM encoder.
 - The encoder performs object detection tasks to detect pre-defined targets such as humans, cars, animals, etc.
 - Using descriptor from the object detection task, the encoder extracts the latent video.
 - The descriptors and extracted latent videos are encoded and multiplexed to be streamed to the VCM decoder as a single bitstream
- b. VCM decoder operation
 - The bitstream is demultiplexed and decoded for further machine tasks

3.7.2 Use case 1: Simple person tracking

- a. Using descriptors, track object class ‘person’.

3.7.3 Use case 2: Specific person tracking

- a. Using descriptors, extract portion from the decoded video containing humans.
- b. Apply additional machine tasks such as gait, height, clothing colour, and hair colour detection to identify the human of interest.
- c. Track the target.

3.7.4 Use case 3: Potential criminal/terrorist tracking

- a. Using descriptors, extract portion from the decoded video containing humans.
- b. Apply additional machine tasks to detect arms and weapons to identify a person in possession of such items.
- c. Track both the person possessing the weapon and the weapon itself in case the weapon gets transferred to another person or the person hides the weapon in a bag or under an overcoat.

3.7.5 Use case 4: Threat alert

- a. While tracking a human of interest from the use case 3, apply action recognition task.
- b. If the person possessing the weapon tries to use the weapon in possession, alert the area.

4 Requirements

- a) VCM shall support video coding for machine task consumption purposes.
Description: The specification shall enable bitstream to be decompressed to perform a machine task of various kinds.
- b) VCM shall support feature coding. (Feature coding)
Description: VCM shall support the input of feature map.
- c) VCM shall support a coding efficiency improvement for at least 30% BD-rate over the VVC standard on machine vision tasks. (video coding)
- d) VCM shall support a broad spectrum of encoding rates.
- e) VCM shall support various degrees of delay configuration.
Description: Ultra-low latency of less than one picture interval could require encoding and decoding operation without picture re-ordering and at units less than a picture. Streaming applications or similar could enable hierarchical coding of pictures and hence incur picture re-ordering delay in encoding and decoding. (Video coding, need further discussion on feature coding)
- f) VCM shall be agnostic to network models. (Video/Feature coding)
- g) VCM shall be agnostic to machine task types. (Video/Feature coding)
- h) VCM shall provide description of the meaning or the recommended way of using the decoded data. (Feature coding)
- i) VCM should support the use and inclusion of information such as descriptors in its bitstream. (Feature/Video coding)
- j) A single VCM bitstream shall support any number of instances of machine tasks. (video coding/feature coding)
- k) VCM shall support at least the following colour formats; monochrome, RGB, and YUV (YCbCr). (Video coding)
- l) VCM shall support at least the following input bit depths: 8-bit and 10-bit. (Video coding)
Note: Other bit depth supports are not prohibited. YCbCr color spaces with 4:0:0, 4:2:0 and 4:4:4 sampling, 8 and 10 bits per component shall be supported. RGB with 4:4:4 sampling, 8 and 10 bits per component shall be supported. BT709 and BT2100 color gamuts shall be supported.
- m) VCM complexity shall allow for feasible implementation within the constraints of the available technology at the expected time of usage.
- n) VCM shall support rectangular picture format up to 7680x4320 pixels (8K).

- o) VCM shall support fixed and variable rational frame rates for video inputs.
- p) VCM shall support any input source from video or image.

Note: For example, the source content may be camera-captured or may have text or graphics overlaid onto a camera-captured scene.

- q) VCM shall support privacy and security (Mandated by ISO)

ANNEX A ADDITIONAL USE CASES

The following machine vision and hybrid machine-human vision use cases can benefit from a system that uses VCM for compression.

A.1. Machine vision use case list

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 or 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

A.2. Hybrid human and machine vision use case list

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

- Timing, which allows cross-modal events to be synchronized and latency accounted for.
- Expressivity, which includes changing haptic parameters based on real-time input.

Haptic Playback for Immersive Media

User expectation for haptic playback is influenced by two main elements: the viewer's location in the scene and viewing orientation. Depending on these elements, haptic effects that are contingent on these two variables may be modulated, mixed, or both. Modulation refers to changing effect parameters, such as intensity or frequency. Mixing refers to combining multiple haptic effects into a single actuator signal that preserves the design intent for the role of haptics in the scene.