|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding  by Artificial Intelligence  www.mpai.community |

|  |  |
| --- | --- |
| **Public Document** | |
| **N378** | 2021/09/30 |
| **Source** | Leonardo Chiariglione |
| **Title** | MMC-HCI Use Case and Functional Requirements WD0.2 |
| **Target** | MPAI-12 |

[Abstract 1](#_Toc83553756)

[1 Introduction 2](#_Toc83553757)

[2 Scope of the Use Cases 3](#_Toc83553758)

[3 Terms and Definitions 4](#_Toc83553759)

[4 References 5](#_Toc83553760)

[4.1 Normative References 5](#_Toc83553761)

[4.2 Informative References 5](#_Toc83553762)

[5 Use Cases 5](#_Toc83553763)

[5.1 Use Case #6 – The Human-CAV Interaction AIW (HCI) 5](#_Toc83553764)

[5.1.1 Use Case description 5](#_Toc83553765)

[5.1.2 Human-CAV Interaction Reference Model 6](#_Toc83553766)

[5.1.3 Input and output data 7](#_Toc83553767)

[6 Functional Requirements 7](#_Toc83553768)

[6.1 Human-CAV Interaction 7](#_Toc83553769)

[6.1.1 I/O Data summary 8](#_Toc83553770)

[6.1.2 Audio 9](#_Toc83553771)

[6.1.3 Autonomous Motion Subsystem Response 9](#_Toc83553772)

[6.1.4 Concept (Face) 9](#_Toc83553773)

[6.1.5 Concept (Speech) 10](#_Toc83553774)

[6.1.6 Emotion 10](#_Toc83553775)

[6.1.7 Face identity 10](#_Toc83553776)

[6.1.8 Face Objects 10](#_Toc83553777)

[6.1.9 Full World Representation 10](#_Toc83553778)

[6.1.10 Full World Representation commands 10](#_Toc83553779)

[6.1.11 Human Commands 11](#_Toc83553780)

[6.1.12 Intention 11](#_Toc83553781)

[6.1.13 Meaning 11](#_Toc83553782)

[6.1.14 Object Identifier 11](#_Toc83553783)

[6.1.15 Speaker Identity 11](#_Toc83553784)

[6.1.16 Text 11](#_Toc83553785)

[6.1.17 Video 12](#_Toc83553786)

# Abstract

*Connected Autonomous Vehicles* (CAV) is one of the 10 current MPAI standard projects. One of them is Multimodal Conversation (MPAI-MMC), a standard comprising the Conver­sation with Emotion and Multimodal Question Answering Use Cases and 3 Speech Translation Use Cases. For the purpose of standardisation, a CAV has been subdivided in 5 subsystems, one of then called Human-CAV Interaction (HCI) whose function is to deal with the technology-rich scenario of the ways a human and a CAV interact.

MMC-HCI is the 6th MPAI-MMC Use Case.

# Introduction

Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) is an [international Standards Developing Organisation](http://mpai.community/) with the mission to develop *AI-enabled data coding standards*. Research has shown that data coding with AI-based technologies is generally *more efficient* than with existing technologies. Compression and feature-based description are notable examples of coding.

In the following, Terms beginning with a capital letter are defined in *Table 3* if they are specific to MPAI-MCS Standard and to **Error! Reference source not found.** if they are common to all MPAI Standards.

MPAI Application Standards enable the development of AI-based products, applications and services. The MPAI AI Framework (AIF) Standard (MPA-AIF) [**Error! Reference source not found.**] provides the foundation on which the technologies defined by MPAI Application Standards operate.

**Error! Reference source not found.** depicts the MPAI-AIF Reference Model. MPAI-AIF provides the foundation on which Implementations of MPAI Application Stan­dards operate.

An AIF Implementation allows execution of AI Workflows (AIW), composed by basic processing elements called AI Modules (AIM).

Diagram

Description automatically generated

*Figure 1 – The AI Framework (AIF) Reference Model and its Components*

MPAI Application Standards normatively specify Semantics and Format of the input and output data and the Function of the AIW and the AIMs, and the Connections between and among the AIMs of an AIW.

In particular, an AIM is defined by its Function and Data, but not by its internal architecture, which may be based on AI or data processing, and implemented in software, hardware or hybrid software and hardware technologies.

MPAI defines Interoperability as the ability to replace an AIF, an AIW or an AIM Implementation with a functionally equivalent Implementation. MPAI also defines 3 Interoperability Levels of an AIF that executes an AIW that is:

1. Proprietary AIW and is composed by AIM with proprietary functions using any proprietary data Format (*Level 1*).
2. Composed of AIMs having all their Functions, Formats and Connections specified by an MPAI Application Standard (*Level 2*).
3. Level 2 and with AIMs certified by an MPAI-appointed Assessor to holdthe attributes of Rel­iability, Robustness, Replicability and Fairness – collectively called Performance (*Level 3*).

MPAI is the root of trust of the MPAI Ecosystem [**Error! Reference source not found.**] offering Users access to the promised benefits of AI with a guarantee of increased transparency, trust and reliability as the Interoperability Level of an Implementation moves from 1 to 3.

# Scope of the Use Cases

A Connected Autonomous Vehicle (CAV) is a mechanical system capable of executing a command to move its body guided by an analysis of data produced by a range of sensors exploring the environment and information transmitted by other sources in range, e.g., CAVs and roadside units (RSU).

MPAI has subdivided a CAV in 5 main subsystems:

*Human-CAV interaction (HCI)*

Recognises the human CAV rights holder

Responds to humans’ commands and queries

Provides extended environment representation (called Full World Repres­entation) for humans to enjoy

Senses human activities during the travel

Activates other subsystems as required by humans or as deemed necessary by the identified conditions.

*Environment Sensing Subsystem (EDS)*

Acquires information from the physical environment via a variety of sensors

Develops the best environment representation (called Basic World Representation).

*Autonomous Motion Subsystem (AMS)*

Computes the Route to destination

Uses dif­ferent sources of information – CAV sensors, other CAVs and transmitting units – to produce a Full World Representation

Gives command that drive the CAV to the intended destination.

*CAV to Everything Subsystem (V2X)*

Teceives information from external sources, including other CAVs, other vehicles and Roadside Units (RSU).

Sends information to other CAVs.

*Motion Actuation Subsystem (MAS)*

Provides environment information

Actuates motion commands in the environment.

The interaction of the 5 subsystems in depicted in *Figure 2*.

Diagram

Description automatically generated

*Figure 2 – The CAV subsystems*

The following high-level workflow illustrates the CAV operation envisaged by this docum­ent.

A *human* with appropriate credentials requests the CAV, via Human-CAV Interaction, to take the human to a given place.

*Human-CAV Interaction* authenticates the human, interprets the request and passes a command to the *Autonomous Motion Subsystem*.

*Autonomous Motion Subsystem*:

Requests *Environment Sensing Subsystem* to provide the current Pose.

Computes the Route.

Issues the start command.

*Environment Sensing Subsystem* computes and sends Basic World Representations to the *Autonomous Motion Subsystem.*

*CAV to Everything*

Becomes aware of other CAVs and external sources (CAVs, RSU etc.).

Shares the CAV’s Basic World Representation with CAVs in range.

*Autonomous Motion Subsystem*:

Receives and processes data broadcasted by external sources (CAVs, RSU etc.).

Computes the Full World Representation.

Shares the CAV’s Full World Representation with CAVs in range.

Computes a Path.

Issues commands to the *Motion Actuation Subsystem* to move the CAV accordingly.

While the CAV moves, *Humans*

Hold conversation with *Human-CAV Interaction* and possibly other hum­ans on board.

Issue commands.

Request views of the environment (Full World Representation) etc.

Interact with other CAVs.

# Terms and Definitions

The terms used in this document whose first letter is capital have the meaning defined in *Table 3*.

*Table 1 – Table of terms and definitions*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Basic World Representation | A description of the Environment based on the CAV sensors and an Offline Map. |
| Command | High-level instructions whose execution allows a CAV to reach a Goal. |
| Connected Autonomous Vehicle | A vehicle capable to autonomously reach an assigned target by understan­ding human utterances, planning a route, sensing and interpreting the environment, exchanging information with other CAVs and acting on the CAV’s motion subsystem. |
| Decision horizon | The estimated time between the current state and the Goal. |
| Emotion | An attribute that indicates an emotion out of a finite set of Emotions |
| Emotion Grade | The intensity of an Emotion |
| Emotion Recognition | An AIM that decides the final Emotion out of Emotions from different sources |
| Environment | The portion of the world within the Decision horizon. |
| Full World Representation | A representation of the Environment using the CAV’s and other CAVs’ Basic World Representations. |
| Goal | The planned State at a future time. |
| Intention | Intention is the result of a question analysis that denotes information on the input question. |
| Language Understanding | An AIM that analyses natural language as Text to produce its meaning and emotion included in the text |
| Meaning | Information extracted from the input text such as syntactic and semantic information |
| Offline Map | An offline-created map of a location and associated metadata. |
| Question Analysis | An AIM that analyses the meaning of a question sentence and determines its Intention |
| Question Answering | An AIM that analyses the user’s question and produces a reply based on the user’s Intention |
| Speech Recognition | An AIM that converts speech to Text |
| Speech Synthesis | An AIM that converts Text or concept to speech |
| State | Pose, Velocity and Acceleration of a CAV at a given time. |
| Text | A collection of characters drawn from a finite alphabet |
| Translation | An AIM that converts Text in a source language to Text in a target language |
| Video analysis | An AIM that extracts features from video |
| Way Point | A point 𝑤𝑖 given as a coordinate pair (𝑥𝑖, 𝑦𝑖), in an Offline Map |

# References

## Normative References

MPAI-MMC Call for Technologies, N154, https://mpai.community/standards/mpai-mmc/#CfT

## Informative References

MPAI-CAV Use Cases and Functional Requirements WD0.4, N377\

# Use Cases

## Use Case #6 – The Human-CAV Interaction AIW (HCI)

### Use Case description

The human-CAV interaction is based on the principle that the CAV is impersonated by an avatar, selected by the CAV right-holder, who has the capability to animate head and face and emit speech that inlude features that display as much as possible the features, e.g., emotion, that would be displayed by a human driver.

Examples are:

1. The CAV’s avatar is reactive to the Environment shows, e.g., it shows an angry face because a driver has made an improper motion.
2. The CAV’s avatar is reactive to a Human, e.g., it shows an appropriate face to a human who has made a joke.

Other forms of interaction are:

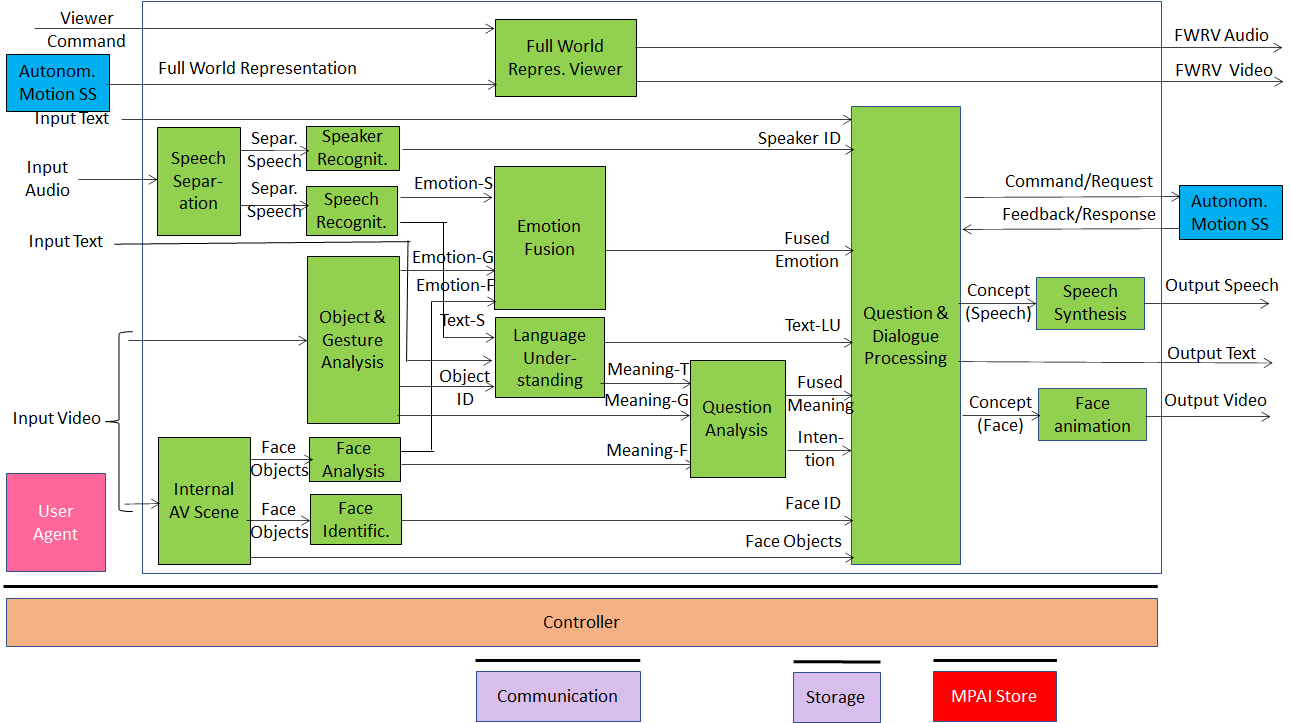
1. CAV authenticates human.
2. A human issues commands to a CAV, e.g.,
   1. Commands to Autonomous Motion Subsystem, e.g.: go to a Way point, display Full World Representation (see 5.3), etc.
   2. Other commands, e.g.: turn off air conditioning, turn on radio, call a person, open window or door, search for information etc.
3. A human entertains a dialogue with a CAV, e.g.,
   1. Information requests, e.g.: time to destination, route conditions, weather at destination etc.
   2. Casual conversation.
4. A CAV monitors the passenger compartment, e.g.,
   1. Physical conditions, e.g.: temperature level, media being played, sound level, noise level, anomalous noise, etc.
   2. Passenger data, e.g.: number of passengers, ID, estimated age, destination of passengers.
   3. Passenger activity, e.g.: level of passenger activity, level of passenger-generated sound, level of passenger movement, emotion on face of passengers.
   4. Passenger-to-passenger dialogue, two passengers shake hands, or passengers hold everyday conversation.

The Human-CAV Interaction collects a variety of data generated by humans inside the vehicle for possible action. This issue if part of the more general problem of data privacy in a CAV that is handled in a dedicated chapter of the Connected Automotive Vehicle document [2].

### Human-CAV Interaction Reference Model

*Figure 3* is the of Human-CAV Interaction (HCI) reference model. The following is noted:

1. A combination of Conversation with Emotion and Multimodal Question Answering AIMs with gesture recognition capabilities covers most Human-CAV Interaction needs.
2. New requirements to existing data format are added where required.
3. Additional AIMs are added when new HCI data formats are required.



*Figure 3 – Human-CAV Interaction Reference Model*

Depending on the technology used (data processing or AI), the AIMs in *Figure 3* may need to access external information, such as Knowledge Bases, to perform their functions. While not represented in *Figure 3*, they will be identified, if required, in the AI Modules subsection.

### Input and output data

*Table 2* lists the input and output data of the HCI AIMs Interaction depicted in *Figure 3*.

*Table 2 – HCI input and output data*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Audio | User outdoor | User authentication  User command |
| Text | User outdoor | User authentication  User command |
| Text | Passenger compartment | Social life of user  Commands or interaction with CAV |
| Audio | Passenger compartment | Social life of user  Commands or interaction with CAV |
| Video | Passenger compartment | Social life of user  Commands or interaction with CAV |
| Full World Representation | Autonomous Motion SS | For processing by FWR Viewer |
| **Output data** | **To** | **Comments** |
| Text | Autonomous Motion Subsystem | Commands to be executed |
| Synthetic speech | Passenger compartment | CAV’s response to passengers |
| Synthetic face | Passenger compartment | CAV’s response to passengers |
| Full World Representation | Passenger compartment | For passengers to view external world |

# Functional Requirements

The Functional Requirements refer to the individual technologies identified as necessary to implement MPAI-CAV Use Cases belonging using AIMs operating in an MPAI-AIF AI Framework. The Functional Requirements developed in this document adhere to the following guidelines:

AIMs are defined to allow implementations by multiple technologies (AI, ML, DP)

DP-based AIMs need interfaces, e.g., to a Know­ledge Base. AI-based AIM will typically require a learning process, however, support for this process is not included in the document. MPAI may develop further requirements covering that process in a future document.

AIMs can be aggregated in larger AIMs. Consequently, some data flows of aggregated AIMs may no longer be accessible.

## Human-CAV Interaction

*Table 3* lists the AIMs represented in *Figure 3*.

*Table 3* *– AI Modules of* *Human-CAV interaction*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Speech detection and separation** | 1. Separates relevant speech vs non-speech signals 2. Detects request for dialogue. |
| **Speaker identification** | Recognises speaker. |
| **Speech recognition** | 1. Analyses the speech input 2. Generates text and emotion output. |
| **Object and gesture analysis** | 1. Analyses video to identify object 2. Produces the ID of the object in focus 3. Analyses video 4. Produces motion and mean­ing of gesture. |
| **Face recognition** | 1. Analyses the video of the face of a human 2. Recognise the human’s identity. |
| **Face analysis** | 1. Analyses the video of the face of a human 2. Extracts emotion and meaning. |
| **Language understanding** | 1. Uses a language model (embedded in AIM) 2. Analyses natural language expressed as text 3. Produces the meaning of the text. 4. Produces text related to Object ID |
| **Emotion recognition** | 1. Fuses Emotions from Speech, Face and Gesture. 2. Produces Final Emotion. |
| **Question analysis** | 1. Fuses Meanings of Speech, Face and Gesture 2. Analyses the meaning of the sentence 3. Determines the Intention. 4. Outputs Final Meaning |
| **Question & dialog processing** | 1. Receives Speaker ID and Face ID 2. If speaker ID and face ID match, then    1. Produces a command to Autonomous Motion SS    2. Analyses user’s emotion, intention, meaning and/ or ques­tion, text    3. Produces Reply (speech) and Reply (face). 3. Else, responds appropriately |
| **Speech synthesis** | Converts Reply (speech) to speech. |
| **Face animation** | Converts Reply (face) to animated face |
| **Full World Representation Viewer** | 1. Receives Full World Representation 2. Presents a view as instructed by human via Commands. |

### I/O Data summary

For each AIM (1st column), *Table 4* gives the input (2nd column) and the output data (3rd column).

*Table 4 – I/O data of Human-CAV Interaction AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Speech Separation** | Input Audio | Separated Speech |
| **Internal AV Scene** | Input Video | Face Objects |
| **Speaker Recognition** | Separated Speech | Speaker ID |
| **Speech recognition** | Separated Speech | Emotion (Speech)  Text (Speech) |
| **Object and Gesture Analysis** | Input Video | Object ID  Emotion (Gesture)  Meaning (Gesture) |
| **Face Analysis** | Face Objects | Emotion (Face)  Meaning (Face) |
| **Face Identification** | Face Objects | Face ID |
| **Full World Representation Viewer** | Viewer Command  Full World Representation | FWRV Audio  FWRV Video |
| **Emotion Fusion** | Emotion (Speech)  Emotion (Face)  Emotion (Gesture) | Fused Emotion |
| **Language Understanding** | Text (Speech)  Input Text  Object ID | Text (Language Understanding)  Meaning (Text) |
| **Question analysis** | Meaning (Text)  Meaning (Gesture)  Meaning (Face) | Fused Meaning  Intention |
| **Question and dialogue processing** | Input Text  Speaker ID  Fused emotion  Text (Speech)  Fused Meaning  Intention  Face ID  Face Objects | Command/Request  Feedback/Response  Concept (Speech)  Output Text  Concept (Face) |
| **Speech synthesis** | Concept (Speech) | Output Speech |
| **Face animation** | Concept (face) | Output Video |

MPAI has acquired a set of first-generation technologies related to the data types listed below in a previous Call for Technologies for the MPAI-MMC standard [1]. MPAI is ready, however, to consider new technologies related to the data below to support new requirements and/or to enhance capabilities.

### Audio

Audio is sampled from an analogue source (passenger compartment) at a frequency in the 44.1-96 kHz range with at least 16 and at most 24 bits/sample.

**To respondents**

Respondents are invited to comment on this choice.

### Autonomous Motion Subsystem Response

The responses of the Autonomous Motion Subsystem are:

1. Enumeration of possible routes with major features of each route.
2. Enumeration of possible parking places with major features of each place.
3. Announcement of obstacles preventing the expeditious accomplishment of the Command.
4. Announcement that the desired Waypoint has been reached.

**To respondents**

Respondents are requested to propose a coded representation of the above commands. Proposals of coded representation of additional responses are welcome.

### Concept (Face)

MPAI-MMC has adopted a Lips Animation format for its MPAI-MMC Standard [1].

**To Respondents**

MPAI is now looking for a technology that can animate head and face of the avatar with the purpose to represent:

1. Motion of head when speaking.
2. Motion of face muscles and eyeballs.
3. Turning of gaze to a particular person.
4. Emotion of the associated spoken sentence.
5. Meaning of the associated spoken sentence.

### Concept (Speech)

MPAI-MMC has adopted Text With Emotion as Reply (speech) format for its MPAI-MMC Standard [1].

**To Respondents**

Respondents are requested to propose a “Concept to Speech” format with the following requir­ements:

1. Capability to represent varying Emotions in the synthetic Speech.
2. Capability to represent varying Meanings of the CAV reply.

### Emotion

MPAI has defined an extensible 3-level set of Emotions for its MPAI-MMC Standard [1].

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [1] for CAV purposes.

### Face identity

The Face Identity AIM shall be able to represent the identity of a limited number of faces.

**To respondents**

Respondents are requested to propose a face identification system suitable for a limited number of faces.

Proposals of a face identification usable in the context of a company renting CAVs to customers are welcome.

### Face Objects

In order for the HCI Subsystem to have a full understanding of what is happening in the passenger compartment (e.g., to have a more natural audio-visual interaction with the passengers, recording of what happened in the compartment etc.), the HCI Subsystem needs to represent the data acqu­ired from the compartment. The current use is

1. To extract the face of a passenger for the purpose of extracting Emotion and Identity.
2. To determine the exact location of a passenger in the compartment in order to animate the CAV’s Avatar Face in such a way that the Avatar gazes into the eyes of the passenger it is talking to.

**To respondents**

Respondents are invited to propose a format for Face Objects to be used as input to Face Analysis, Face Identification and Question and Dialogue Processing satisfying the above requirements.

### Full World Representation

The requirements of the FWR AIM are developed in the context of CAV Autonomous Motion Subsystem requirements.

### Full World Representation commands

The requirements of FWR interaction will be developed once the FWR requirements are defined.

### Human Commands

The basic commands given to the Autonomous Motion Subsystem are:

1. Go to a Waypoint.
2. Park close to a Waypoint.
3. Drive faster.
4. Drive slowly.
5. Display Full World Representation.

**To respondents**

Respondents are requested to propose a coded representation of the above commands. Proposals of coded representation of additional commands are welcome.

### Intention

MPAI has defined a digital representation format for Intention in [1].

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [1] for CAV purposes.

### Meaning

MPAI has defined a digital representation format for Meaning [1].

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [1] for CAV purposes.

### Object Identifier

MPAI has defined a digital representation format for Object Identifier in [1].

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [1] for CAV purposes.

### Speaker Identity

The current Speaker Identity requirements demand the ability to identify a limited number of Speakers.

**To respondents**

Respondents are requested to propose a Speaker Identification methods suitable for a limited number of speakers.

Proposals of a Speaker Identification method usable in a content of a company renting CAVs to customers are welcome.

### Text

As there is a need to support most languages in use, Text representation conforms to ISO/IEC 10646, Information technology – Universal Coded Character Set (UCS).

**To respondents**

Respondents are invited to comment on this choice.

### Video

Video is intended for use in the passenger compartment.

The following characteristics of 2D Video have been adopted.

1. Pixel shape: square
2. Bit depth: 8-10 bits/pixel
3. Aspect ratio: 4/3 and 16/9
4. 640 < # of horizontal pixels < 1920
5. 480 < # of vertical pixels < 1080
6. Frame frequency 50-120 Hz
7. Scanning: progressive
8. Colorimetry: ITU-R BT709 and BT2020
9. Colour format: RGB and YUV
10. Compression: uncompressed, if compressed AVC, EVC, HEVC.

**To respondents**

Respondents are invited to comment on MPAI’s choice for 2D Video.

Respondents are also requested to propose a data format for 3D Video having video+depth as the baseline format or other 3D Video data formats.