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

|  |  |
| --- | --- |
| N68 | 2020/11/18 |
| Source | MPAI-EVC Group |
| Editors | Ajay Luthra, Leonardo Chiariglione |
| Title | Draft MPAI-EVC Use Cases and Requirements |
| Target | MPAI Video |

**ABSTRACT**

This document describes various applications that use digital video coding technologies & standards, and their operating environments and attributes. It also describes the requirements that the next generation of video coding standard should meet to satisfy the needs of those applications.

# INTRODUCTION

Innovations in the areas of Artificial Intelligence (AI) and Machine Learning (ML) technologies and their implementations have been increasing exponentially during last few years. Their usage can now be found in wide range of applications and areas, including Image Processing, Image Recognition, eCommerce, Workplace Communication, Healthcare, Agriculture, Cybersecurity, Finance, Autonomous Vehicles, Supply Chain Management, Manufacturing etc.

MPAI aims to use the advances in AI technologies to code multiple data types more efficiently and to develop standards for data coding that have Artificial Intelligence (AI) as its core technology [1]. One of those data coding standard areas is called MPAI AI-Enhanced Video Coding (MPAI-EVC). It is focused on understanding how AI/ML can help in improving the performance of existing digital video coding technologies and standards. It is expected that if AI/ML based technologies are found to help in significantly improving the performance over state-of-the-art existing video codecs, an MPAI Video coding standard will be developed.

Section 2 describes various Use Cases that use digital video coding technologies & standards, and their operating environments and attributes. Section 3 describes the requirements that the next generation of video coding standard should meet to satisfy the needs of those Use Cases.

It is a work in progress. This topic was first discussed during the MPAI-EVC’s teleconference on Oct 13, 2020. It is expected to be further refined and developed during future meetings as well as via email discussions.

# Use Cases / Applications

## Entertainment TV Content Distribution

This application includes sending entertainment content to home. It is sometimes also referred as “Watch TV’ application. Content can be viewed on a large screen TV as well as small screens of mobile devices. Complexity of the encoders can be significantly higher than that of decoders. Content gets compressed in two modes:

* Real time encoding
* Off line encoding

Real time encoding is used for live contents like sports, news etc. Off-line encoding is used for stored content for application like On-Demand TV, Streaming Video etc. Compressed content can also be stored either in the network for nPVR (networked Personal Video Recoding) application or in-home DVR (Digital Video Recorder), including whole home DVR. In-home DVR can also distribute the stored compressed content to other devices in home, including other TVs and mobile devices.

Transcoding of the content may also be done either to change the bit rates or to use another encoding standard that can be decoded by a receiving device. Transcoding functionality can be implemented in IRDs (Integrated Receiver Decoder), commercial inserting devices or advanced nPVRs/DVRs.

Commercials may also be inserted in some content. Commercial inserters may also perform transcoding and/or transrating to match the coding standard and/or the bit rate to that of the main stream.

In summary, following are the primary attributes of the environment associated with this application:

* Network: Guaranteed QoS, as well as, Best Effort
* Protocols: MPEG-2 TS as well as IP based ABR
* Content Types: Natural video (Camera captured content), Computer generated graphics, Hybrid
* Content Attributes
  + HD, 4k, 8k
  + Aspect ratio: 16:9 is the most popular
  + SDR and HDR
  + Standard and Wide Color Gamut
  + 4:2:0 Color format
  + Bit Depth: Mainly, up to 10 bits.
  + Frame Rates: up to 120 Hz
  + Some content may include film noise
  + Transition points between two video sequences may include blending of the video
* Compression Mode: Real Time and Off-line
* Compression type: Lossy
* End-to-end delay: Mainly, High Delay (> 100 ms to off-line encoding)
* Implementation architecture
  + Decoders: Software as well as Hardware/ASIC based
  + Encoders
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
    - Significantly more complex than decoders
    - Some encoders may include film grain pressing modules
  + Transcoders
    - Decode and re-encode the content to convert the compression standard and/or bit rates
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
* Viewing Environment: 2D
* Accessibility: To be able to do random access with 1 to 6 sec delay. Picture level accessibility is not necessary and long Group of Pictures (GOP) structures can be used.
* Special Effects: To be able to pause, reverse play back, fast forward, slow play back etc.
* Commercial Insertion: Commercials, including the ones that may not have the same attributes as the main video, may be inserted in the compressed domain.

## Video Games, including eSports and Cloud Gaming

From the original game machines, such as Play Station and XBox, the market has extended to also include online/cloud games:

* In traditional online games the logic of the game is executed by a server that receives commands from clients, processes them and sends the processed commands to the clients which displays locally the appropriate video frame.
* In cloud gaming commands are sent to a virtual machine on the cloud that runs the logic of the game and may send further commands to a server in case of a multiuser game. The virtual machine creates and encodes the video frame to the remote client. The virtual machine has a lot more information about the video that a regular video encoder has.
* In eSports (electronic sports), multiple players play the game against each other with many spectators watching the game on the network through live streaming.
* Online gambling consists of gambling conducted on the internet. This includes virtual games (like, poker, bingo etc.) and sports betting etc. Sports betting includes predicting sports (e.g. football, basketball, baseball, hockey, cycling, auto racing, boxing etc.) results and placing a wager on the outcome. In this application video is streamed from one of multiple points to one or multiple points. The video could include the video of the sport or virtual games being played. Virtual games can also include graphics (computer generated) information.
* MPAI is exploring Server-based Predictive Multiuser Gaming [1] where an AI-machine uses available context information (e.g. other users, past games) to make up for the information that has not reached the server. A standard interface between MPAI-SPG and the MPAI-EVC video encoder would be beneficial. In this use case, objects in the scene are known to the encoder. Object-based coding could also be used to encode various objects with a known background. Those objects could be reconstructed at the decoder and overlaid on that known background.

Following are the primary attributes of the environment associated with this application:

* Network: Best Effort
* Protocols: IP based ABR
* Content Types: Computer generated graphics, Hybrid
* Content Attributes
  + HD, 4k, 8k
  + Aspect ratio: 16:9 is the most popular
  + SDR and HDR
  + Standard and Wide Color Gamut
  + 4:2:0 Color format
  + Bit Depth: Mainly, up to 10 bits.
  + Frame Rates: up to 120 Hz
* Compression Mode: Real Time
* Compression type: Lossy
* End-to-end delay: Low (30 msec < Delay <100 msec) to Very Low Delay (< 30 msec; less than one picture period)
* Implementation architecture
  + Decoders:
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
  + Encoders
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
    - Comment: Encoder has information on how (synthetic) objects in the scene move.
* Viewing Environment: 2D, including Head Mounted Devices (HMD)

## Videoconferencing

With the current pandemic, videoconferencing is playing a very important role in private and professional life presenting new challenges for a technology that was largely conceived for professional use in professional environments. Many are using videoconferencing from their homes where they would like the system to automatically hide components of the scene which are considered unsuitable.

Following are the primary attributes of the environment associated with this application:

* Network: Best Effort
* Protocols: IP based ABR, RTP
* Content Types: Natural video (Camera captured content), Computer generated graphics, Hybrid
* Content Attributes
  + Mainly HD, sub-HD
  + Aspect ratio: 16:9 is the most popular for rectangular video
  + May include object-based representation, (including background-foreground representation)
  + Mainly SDR
  + Mainly Standard Color Gamut
  + 4:2:0 Color format
  + Bit Depth: Mainly, up to 10 bits.
  + Frame Rates: up to 60 Hz
* Compression Mode: Real Time
* Compression type: Lossy
* End-to-end delay: Low
* Implementation architecture
  + Decoders: Software as well as Hardware/ASIC based
  + Encoders:
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
    - Significantly more complex than decoders
* Viewing Environment: 2D

## Social Media

## Video content is increasingly becoming more and common among social media engagements and traffic on applications like Facebook, LinkedIn, YouTube, TikTok, WhatsApp. These applications include a large amount of user/consumer generated content, including various How-to-do video content. Content can be both generated by and consumed on mobile devices.

Following are the primary attributes of the environment associated with this application:

* Network: Best Effort
* Protocols: IP based ABR
* Content Types: Natural video (Camera captured content), Computer generated graphics, Hybrid
* Content Attributes
  + Mainly sub-HD, HD
  + Aspect ratio:
    - Main Video –Various
    - Video banners – 2:1, 6:1, 8:1,
    - Vertical video in social media - 4:5, 2:3, and 9:16
  + Mainly SDR
  + Mainly Standard Color Gamut
  + 4:2:0 Color format
  + Bit Depth: Mainly, up to 10 bits.
  + Frame Rates: up to 60 Hz
* Compression Mode: Real Time and Off-line
* Compression type: Lossy
* End-to-end delay: High
* Implementation architecture
  + Decoders: Mainly software based.
  + Encoders:
    - Mainly software based
    - Stand alone
* Viewing Environment: 2D

## Drones Based / Wearable Cameras

Digital imaging technology, miniaturized computers, and numerous other technological advances over the past decade have contributed to rapid increase in use of Drones has increased steadily during last few years for visual imagery. Applications include capturing video of social events (like, weddings, large parties etc.), security, military surveillance, agricultural imagery, traffic monitoring, electronic news gathering, remote sensing etc. Popularity of wearable cameras is also increasing in consumer, military and policing applications.

Following are the primary attributes of the environment associated with this application:

* Network: Best Effort as well as Guaranteed QoS
* Protocols: MPEG2 TS, IP based ABR, RTP
* Content Types: Natural video (Camera captured content), Infrared images, Monochrome video
* Content Attributes
  + Mainly HD, sub-HD. Migration to 4k is expected in future.
  + Aspect ratio: 4:3 and 16:9 ratios are the most popular for rectangular video
  + Mainly SDR
  + Mainly Standard Color Gamut
  + 4:2:0 and 4:0:0 Color formats. Some applications migrating to 4:2:2.
  + Bit Depth: Mainly, up to 10 bits.
  + Frame Rates: up to 60 Hz
* Compression Mode: Real Time
* Compression type: Lossy
* End-to-end delay: High. Some applications require Low delay.
* Implementation architecture
  + Decoders: Software as well as Hardware/ASIC based
  + Encoders:
    - Hardware/ASIC based
    - Stand alone
    - Low power and small size
* Viewing Environment: 2D

## Medical Video

Coding of high-resolution CAT scans to map the brain, MRI scans, Ultrasound, Endoscopic images for the purposes of storing it and/or communicating to the viewers.

In summary, following are the primary attributes of the environment associated with this application:

* Network: Guaranteed QoS as well as Best Effort
* Protocols: MPEG-2 TS as well as IP based ABR
* Storage: BluRay, Hard Disk, SSD
* File format: ISOBMFF
* Content Types: Captured by non-visual sensors
* Content Attributes
  + SD, HD, 4k, 8k
  + Aspect ratio: 4:3 and 16:9 ratios are the most popular
  + SDR and HDR
  + Standard and Wide Color Gamut
  + Mainly Monochrome, some video content can have 4:2:0 or 4:2:2 Color format.
  + Bit Depth: 10 bits or higher.
  + Frame Rates: Mainly up to 60 Hz
* Compression Mode: Real Time and Off-line
* Compression type: Visually Lossless to Lossy
* End-to-end delay: High. Some applications may require Very Low delay.
* Implementation architecture
  + Decoders: Mainly Software based
  + Encoders
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
    - Significantly more complex than decoders
* Viewing Environment: Mainly 2D. In future immersive environment may be used more widely.
* Accessibility: To be able to do Picture Level accessibility. Long Group of Pictures (GOP) structures can also be used.

## Telemedicine

This application includes remote diagnosis and video conferencing type patient-doctor meetings.

Following are the primary attributes of the environment associated with this application:

* Network: Best Effort
* Protocols: IP based ABR
* Content Types: Natural video (Camera captured content), Computer generated graphics, Hybrid
* Content Attributes
  + Mainly HD, sub-HD
  + Aspect ratio: 16:9 is the most popular for rectangular video
  + Mainly SDR
  + Mainly Standard Color Gamut
  + 4:2:0 Color format
  + Bit Depth: Mainly, up to 10 bits.
  + Frame Rates: up to 60 Hz
* Compression Mode: Real Time
* Compression type: Lossy
* End-to-end delay: Low
* Implementation architecture
  + Decoders: Mainly software based
  + Encoders:
    - Mainly software based
    - Stand alone
* Viewing Environment: 2D

## Security / Surveillance

Video surveillance use cases involve monitoring of an area and/or traffic via video cameras. Those cameras can either be land based or drone based or wearable. Those cameras are connected to a recording device or IP network, and may be monitored visually.

Use case and associated requirements of the drone-based or wearable cameras is covered in another section. Focus in this section is on the ground based capturing devices.

If AI based MPAI-EVC architectures can help in analyzing footage, organizing digital video footage into a searchable database, and/or ability to detect fake videos, will helpful in this use case.

Following are the primary attributes of the environment associated with this application:

* Network: Best Effort
* Protocols: IP based ABR
* Content Types: Natural video (Camera captured content), Monochrome video
* Content Attributes
  + Mainly HD, 4k/UHD
  + Aspect ratio: 16:9 is the most popular for rectangular video
  + Mainly SDR
  + Mainly Standard Color Gamut
  + 4:2:0 and 4:0:0 Color formats
  + Bit Depth: Mainly, up to 10 bits.
  + Frame Rates: up to 60 Hz
* Compression Mode: Real Time
* Compression type: Lossy
* End-to-end delay: High, Low
* Implementation architecture
  + Decoders: Software as well as Hardware/ASIC based
  + Encoders:
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
    - Significantly more complex than decoders
* Viewing Environment: 2D

## Digital Cinema

Historically, movies have been distributed and projected using reels of the motion picture film, e.g. 35 mm film. Recently, the movie industry has been migrating towards using digital technology to distribute and project motion pictures. Instead of shipping film reels to movie theaters, a compressed digital version of the content is increasingly getting distributed over either dedicated links or digital storage media.

Following are the primary attributes of the environment associated with this application:

* Network: Typically, dedicated link.
* Content Types: Natural video (Camera captured content), Computer generated graphics, Hybrid. Original content may be captured on a film and converted to a digital format.
* Storage: HDD, SSD, Optical disks.
* Content Attributes
  + HD, 4k, 8k
  + Aspect ratio: Varied - 16:9 to 2.39:1. Other aspect ratios are also used.
  + SDR and HDR
  + Standard and Wide Color Gamut
  + Mainly, 4:4:4 Color formats (some content may use XYZ domain)
  + Bit Depth: 12 to 16 bits.
  + Frame Rates: up to 120 Hz
  + May contain film noise
  + Some content may be in Stereoscopic (3D) or Immersive (e.g. IMAX) format.
* Compression Mode: Mainly, Off-line
* Compression type: Visually lossless to low loss. Compressed bit rates are very high.
* Encoding delay: Not important
* Implementation architecture
  + Decoders: Software as well as Hardware/ASIC based
  + Encoders
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
    - Significantly more complex than decoders
    - Some encoders may include film grain processing modules
  + Transcoders
    - Decode and re-encode the content to convert the compression standard and/or bit rates
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
* Viewing Environment: Mainly 2D. Some movies are captured and displayed in Stereoscopic (3D) formats. Some content may provide immersive experience (e.g. IMAX)
* Accessibility: Typically, Intra-only GOP structures are used.

## Professional Content Creation and Production in Studios

In the creation phase, the content gets captured via high-end cameras, or generated using computers, or converted from the film-based originals. It can also go through multiple stages of processing, including editing, color correction, blending, green/blue screening etc. It is required that the visual quality of the content is very high and remains high as it goes through various production related processing. AI based systems are becoming more commonly used in the content production area also. It is desirable for the MPAI-EVC standard to be able to interface easily with those systems.

Following are the primary attributes of the environment associated with this application:

* Network: Guaranteed QoS
* Protocols: Various in-studio protocols, IP
* Content Types: Natural video (Camera captured content), Computer generated graphics, Hybrid
* Storage: HDD, SSD, Optical disks.
* Content Attributes
  + HD, 4k, 8k
  + Aspect ratio: 16:9 is the most popular. Other aspect ratios are also used.
  + SDR and HDR
  + Standard and Wide Color Gamut
  + 4:2:0, 4:2:2 and 4:4:4 Color formats
  + Bit Depth: 10 to 16 bits.
  + Frame Rates: up to 120 Hz
* Compression Mode: Real Time and Off-line
* Compression type: Visually lossless to low loss (higher bit rates than those normally used for distributing the content to the consumers)
* Encoding delay: Low
* Implementation architecture
  + Decoders: Software as well as Hardware/ASIC based
  + Encoders
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
    - Significantly more complex than decoders
  + Transcoders
    - Decode and re-encode the content to convert the compression standard and/or bit rates
    - Multiple generations (typically less than around 8) of transcoding
    - Software as well as Hardware/ASIC based
    - Stand alone as well as Cloud based
* Viewing Environment: Mainly 2D
* Accessibility: To be able to do random access at Picture level. Typically, short Group of Pictures (GOP) or Intra-only structures are used.

# Requirements

## Coding Efficiency

* Must be able to provide around 25 to 50% bit rate reduction over existing state-of-the-art video coding standard(s) for similar visual quality
* Content Types
* Must be able to provide desired improvement in the coding efficiency for wide range of content types: Natural video (Camera captured content), Computer generated graphics, Hybrid, Video gaming content.

## Content Attributes

* Must be able to provide desired improvement in the coding efficiency for the content with wide range of attributes:
  + Mainly, up to 8k resolution (but should be able to handle higher resolution content)
  + Mainly rectangular video with wide range of aspect ratios, including video banners and vertical video. Some applications, e.g. Videoconferencing, may have object based (including foreground-background) representation
  + SDR and HDR
  + Standard and Wide Color Gamut
  + 4:2:0, 4:2:2 and 4:4:4 Color formats (initial focus is on YUV based coding)
  + Bit depth: Initial focus is on 10 bits video. Should be extensible to higher bit depths.
  + Frame Rates: up to 120 Hz

## Compression Types

* Initial focus is on Lossy compression and visually Lossless compression.
* Mathematically Lossless compression is not the focus at this stage.

## Compression Modes

Standard shall allow the encoders designed for operating in the following modes:

* Real Time
* Off-Line Coding (may also include faster than real time encoding)

## Viewing Environment

* 2D, including HMD
* 3D (Stereoscopic) (Note: not under consideration in Phase 1)
* Immersive (Note: not under consideration in Phase 1)

## Distribution Networks

Digital video content gets distributed via multiple distribution networks, like Cable, Satellite, Telco, Cellular networks (e.g. 5G), Over-the-Air and storage media. From the QoS point of view, the distribution channels provided by these networks fall in two categories:

* Guaranteed channel bandwidth and capacity
  + The available bandwidth in this case is generally known and fixed. It also has low error rate after FEC.
  + Typically, MPEG-2 TS protocol is used.
  + Due to FEC, the error rate at the video layer is small
* Best effort channel capacity
  + Mode of operation here is to provide user the bandwidth that is available at the time of the distribution. The available bandwidth in this case is not known a-priori and is also time-varying.
  + Typically, TCP/IP based Adaptive Bit Rate (ABR) streaming is used.
  + Due to TCP/IP, the packet loss at the video layer is absent.

Standard must be able to support distribution of video over both the network-types above.

* Storage media
  + DVD, BluRay, HDD, Solid State
  + File format: Typically use ISOBMMFF

## Rate Control

## CBR, VBR, Capped VBR

## End-to-end delay

Standard shall be able to support various end-to-end delay configurations:

* High Delay (> 100 ms to off-line encoding)
* Low Delay (30 msec < Delay <100 ms)
* Very Low Delay (< 30 msec; less than one picture period)

## Accessibility

Standard shall allow entering/accessing into the compressed video bit stream with varying accuracy:

* Picture level/resolution (Editing, Splicing etc.)
* Greater than one picture lag/resolution (channel change, commercial insertion etc.)

## Implementation

Friendly to multiple implementation architectures:

* Hardware
* CPU
  + Should be possible to design encoder with architecture that allows tradeoffs among coding speed vs coding efficiency vs number of available CPUs/Cores
* CPU+GPU/FPGA
  + Should be possible to design encoder with architecture that allows tradeoffs among coding speed vs coding efficiency vs number of available CPUs/Cores and presence of GPU and/or FPGA
* CPU+GPU+NPU
  + Should be possible to design encoder with architecture that allows tradeoffs among coding speed vs coding efficiency vs number of available CPUs/Cores and presence of GPU and/or FPGA and/or NPU
* ASICs
* System
  + Stand alone
  + Cloud based
    - Support Virtual Machine architectures
    - Should be possible to design encoder with architecture that allows tradeoffs among coding speed vs coding efficiency vs number of CPUs/Cores available
  + Mobile devices
    - With the exception of a few use cases, like Digital Cinema, Professional Video Creation, in virtually all the Use Cases above, both mobile as well as non-mobile devices can be used today. This imposes the general overall requirement of friendliness to mobile application on MPAI-EVC design and standard. This includes the ability to have designs that are sensitive to the power consumption and, if desired, allow one to build encoders and decoders with architectures that can provide easy trade-off between power consumption and visual quality.
* Complexity
  + In large number of applications, encoder complexity can be significantly higher than the decoder complexity.
    - Off-line encoding applications can tolerate very high asymmetry between encoder-decoder complexity. Those applications may also do multi-pass encoding.
    - Real time encoders have relatively less complexity than that of off-line encoders.
    - In some applications with two-way communication, especially those using mobile devices, it may be more desirable to have less asymmetry between encoder and decoder complexity.

Standard should provide the capability to efficiently trade-off encoder and decoder complexities to match the needs of various applications.

## Special Effects and Editability

Standard shall be able to support:

* Special effects: Pause, Reverse play back, Fast Forward, slow motion play back.
* Editability: concatenation of two video streams in the compressed domain.

## Backward Compatibility/Scalability/Multiple Layer Bitstreams

MPAI-EVC may focus on four types of Scalabilities to achieve some form of backward compatibility:

* Temporal Scalability

In this approach a subset of bit stream can be extracted and decoded to provide a video with lower frame rate. For example, full bit stream may have 100 (or 120) frames/sec frame rate and the subset may provide 50 (or 60) frame/sec

* Signal-to-Noise Ratio (SNR) / Distortion Scalability

In this approach, inner layer of the bit stream provides video with certain distortions (compression artifacts). A decoder can take a greater number of layers to decode video with smaller visual distortion (i.e. improving SNR)

* Spatial Scalability

In this approach, inner layer of the bit stream provides video with certain spatial resolution (for example, 1920x1080). A decoder can take a greater number of layers to decode video with higher spatial resolution (for example, 3840x2160) video.

* Codec Scalability

In this approach, inner layer of the bit stream provides video coded based on certain standard (for example, MPEG-5 EVC) with certain quality (e.g. distortions and/or spatial resolution) and other layers can provide improved quality video. In these methods, a decoder compliant with older video coding standard (e.g. MPEG-5 EVC) is able to decode partial (inner layer) bit stream. This is also sometimes referred as Vertical Hybrid scheme in MPAI-EVC video codec architecture.

## Desirable Features

### Codec Agnostic AI Based Improvements

* It is desirable to design MPAI-EVC such that the AI based coding technologies developed during this standardization process can also be used in improving coding efficiency of wide range to existing video coding standards.

### Pre and Post processing

* If AI based MPAI-EVC architectures can help in analyzing footage, organizing digital video footage into a searchable database, and/or ability to detect fake videos, it will be helpful in various use cases.
* If AI based MPAI-EVC architectures are able to reduce the noise in video more intelligently than blind signal processing, it will be helpful in many use cases.
* If AI based MPAI-EVC architectures are able to detect and process film grain more intelligently than blind signal processing, it will be helpful in compressing and having better visual quality of the content that is originally captured on film.

## Error resilience

## At this stage, MPAI-EVC is not focused on this aspect.

# References

[1] Leonardo Chiariglione etal., “AI-based Data Coding Standardization,” November 2020, MPAI document number M61.