Media Services beyond 2D: Use Cases

Mobile broadband is used today to distribute traditional audiovisual services to smartphones, home routers or connected devices such as set-top boxes (STBs). This includes distribution of live and on-demand streaming services to apps and browsers, experienced in a two dimensional (2D) format with some degree of interactivity provided within them.

This report presents a collection of use cases and high-level architectures, along with an overview of high-level requirements for media services beyond 2D, while subsequent studies might focus on the extent to which 3GPP and other media-related specifications can support these use cases.

The primary scope of the present document is to provide:

1. An overview of the evolution of media services beyond 2D;
2. A collection of existing and future use cases and services;
3. A high-level perspective on architecture, features and requirements

Contents

The evolution of media services beyond 2D

Mobile broadband is used today to distribute traditional audiovisual services to smartphones, home routers or set-top boxes (STBs). This includes distribution of live and on-demand 2D streaming services to apps and browsers, with some degree of interactivity provided within them.

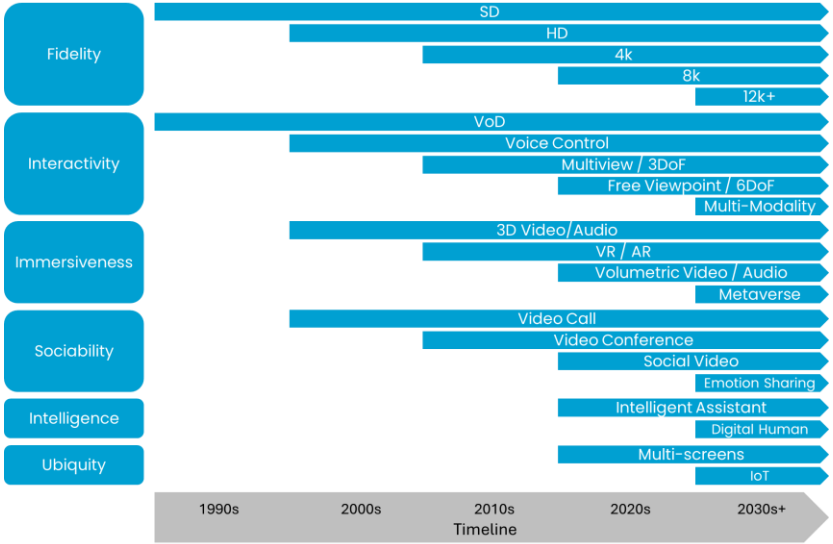
An increasing number of services that expand the user experience beyond 2D formats are under experimentation or becoming part of commercial offers. The implementation of these services may result in specific requirements in terms of capture, contribution, production, encoding, distribution, decoding, rendering, security and user interfaces.

Generally, such services are expected to be developed independently of distribution networks and to rely on IP/cloud/CDN-based delivery systems. Additional relevant attributes include high volumes of data, complex decoding and rendering, and requirements to stay within limits with respect to interactivity and immersivity, such as for motion-to-photon and motion-to-sound latencies. Split-compute and split-rendering distribution architectures are expected to be relevant to make such services feasible and scalable on existing and emerging end devices including TVs, phones, tablets, head-mounted displays (HMDs), AR glasses and other devices.

3GPP has addressed and defined a set of technologies to support several of the above scenarios, at both network and application/service levels. These include, for example, omnidirectional 3DoF VR experiences (Rel-15 and Rel-17), XR and AR supported by split-rendering (Rel-17/18), AV production, etc.

Nevertheless, only a subset of the aforementioned use cases may be fully supported by 3GPP networks and services as of today.

Advancements in connectivity, network technologies, encoding and/or computing have enabled enhancements in user experience, initially with respect to fidelity, but also towards interactivity, immersiveness, sociability, intelligence and ubiquity.



This report focuses on seven categories that may support nine use cases taken as reference in this report to define requirements and underlying architectures and features. These may include, but are not limited to:

Interactive video services: a user has the ability to interact with the produced video and influence or define how the story is told, or even interact with how it is being produced.

Multiview services: a user can select between different camera angles of the same event or consume a multiplicity of such views at the same time.

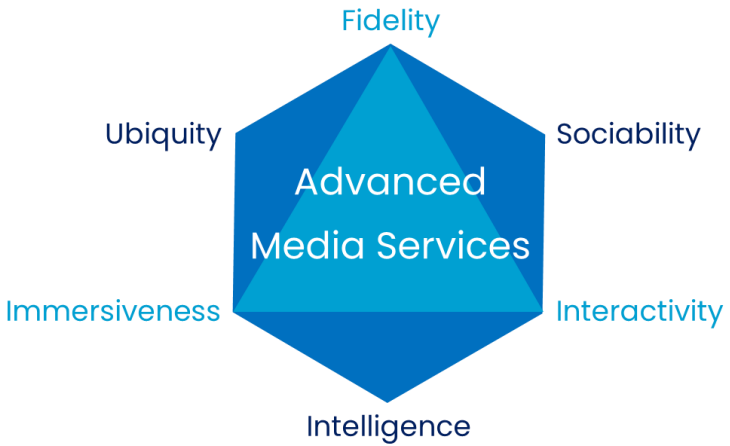
Virtual reality (VR) experiences: a user experiences a service on a VR headset and can have an omnidirectional (e.g., 180° or 360°) experience. Such experiences can be further enhanced, by for example providing multiple viewpoints, mosaics and other viewing enhancements.

Glasses-free 3D media: similar to interactive videos presented, for example, in a HTML5/browser environment, 3D scenes may comprise different types of 2D, volumetric or 3D media assets that can be composed and rendered based on user interaction, location and viewpoint. 3D scenes may be generated by game engines.

Volumetric video: a technique that captures a three-dimensional entity, such as a location or performance, that can be viewed on flat screens, 3D displays and VR HMDs. The viewer generally experiences the result produced in a real-time engine and has the ability to explore the generated volumetric media experience.

Free viewpoint video: free viewpoint video (FVV) offers flexible viewpoint navigation in 3D space and time (referred to as 4D video) from multiview captured video. The user can navigate in six degrees of freedom (6DOF) through the content and control the desired viewing angles and positions.

Social video: participants in social video experiences can interact with each other while watching video in real time, even when located in different places. Interactions could include audio-, video- and text-based chat between the users. Synchronized playout is key to these experiences.



Interactivity: Massively Interactive Live Events

This use case combines the ability to cater for large-scale crowds with audience interactions within the XR experience. It extends a radio show (e.g. a live music show) or popular music festival, typically offered to large audiences on broadcast radio or TV, with an XR experience. This makes it possible to offer more personalized and immersive experiences to audiences than can be achieved through passive listening or viewing alone. This scenario also enables remote participation in big events. The video found on the following page shows the concept behind this use case (starting from 3'31"): https://www.bbc.co.uk/rd/about/vision

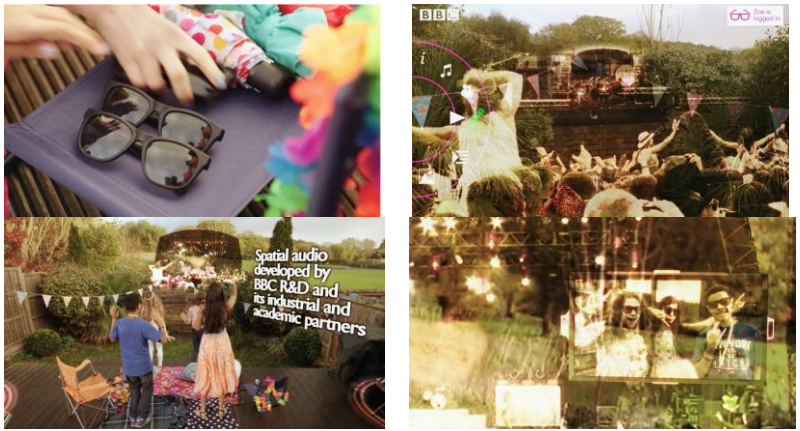


Figure 3: Scenario involving XR devices, a virtual experience blended with a real environment, multiple-user participation and interactions with events.

Architecture, relevant features and pre-conditions

This use case aims to complement live in-person events (bottom left in Fig. 3) also offered on TV and/or radio (top left) with a virtual experience that allows remote audience participants to join from their PCs (or possibly an XR headset), embodied as avatars (bottom right). Performers are captured and reconstructed as photorealistic streams for the virtual experience (top right).

The immersive and personalized experience allows XR participants to be able to respond to the artists in a variety of ways such as emoting their reactions from a range of predefined emotes so they are visible to all attendees in the online world. Virtual participants can also communicate with each other by speaking and listening naturally, taking advantage of proximity-based audio using a microphone and speaker. Despite massive participation, audiences should be provided with an intimate experience.

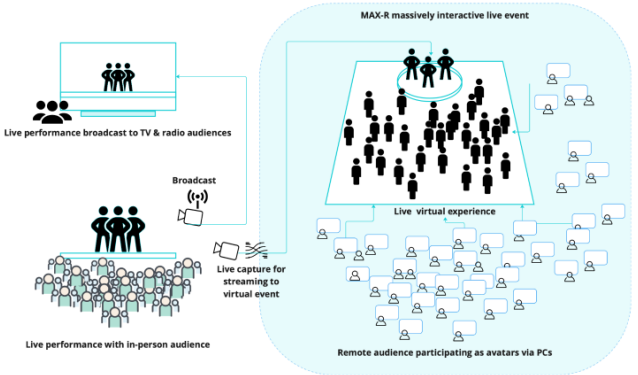


Figure 4: Overview of the concept

High-level perspective on potential requirements

Supporting multiple user environments

Users enjoy the event in a variety of locations and social arrangements: a single user connecting alone or accompanied by other participants in the same physical space, indoors (home, theatre…) or outdoors (park, garden…).

Photorealism and avatars

The virtual event is provided both on a metaverse platform and on TV. Remote participants should see a photorealistic view of the performers (not animated avatars). Users joining the virtual experience should see a photorealistic version of the performance from a limited range of viewpoints, with no need to support viewing from arbitrary locations. The photorealistic view of the performers can be placed in a 3D / game engine / metaverse scene.

Figure 5. Example of 2.5D capture where an actor is recorded, segmented from the background (left), reconstructed in 2.5D (middle), and teleported into a 3D environment, including shadow casting and relighting (right).

This approach complements fully volumetric capture optimized for 360° capture of individuals or smaller areas/volumes. The latter approach tends to be expensive, places constraints on artists and generates large data rates expensive to deliver and can't easily capture large volumes.

Online participants joining the event should be represented by avatars.

QoS/QoE considerations

It should be possible to specify which parts of the performance area are captured for linear viewing and which for the immersive experience.

Latency between the live action and the remote experience should be low enough to allow interactions from audience to reach the performer with minimum delay. Audio should be presented with negligible timing offset compared to video.

Distribution aspects

Visual quality should not be below that which the viewer's device can display.

Participants should be able to join on their personal devices without having to install plug-ins.

The virtual experience should support multiple participants (across the service provider's user population) and should be able to be delivered to multiple users in the same physical location.

XR participants should be able to respond to the artists in a variety of ways such as emoting their reactions visible to all attendees in the online world. Participants should be able to communicate with each other speaking and listening naturally.

Interactivity: Interactive Hub

An interactive media experience centred around the home and other personal spaces, where participants enjoy a multi-sensory experience using smart devices such as mobile phones, TVs, infotainment systems in cars, etc.

Case A: Interactive media (fitness, sports, games, …)

Users can engage in fitness training at any time through intelligent devices with connected cameras. With the help of mobile network edge computing combined with AI capabilities, real-time high-precision human skeleton movement tracking algorithms can be applied for professional fitness guidance. Training results can be professionally rated, analysed, and shared.

Case B: Intelligence assistance and digital human companion

Users can get one-on-one life, work and health assistance and companionship. Users can access these features through smart devices, such as mobile phones, HMDs, smart STBs or TVs with cameras, anytime, anywhere. In addition, edge computing capabilities for media processing and rendering can enable a more immersive interaction with digital humans.

Figure 6. Interactive fitness (left) and intelligence assistance (right)

Case C: Seamless multi-screen transfer

Users can transfer media across multiple screens online and in real time with other viewers (e.g. family or friends).

Figure 7. Seamless multi-screen transfer

Architecture, relevant features and pre-conditions

These are examples of typical deployment scenarios for an Interactive Hub. In general, users can access the services via smart devices, with screens and cameras, connected to the network. When the computing power on the device is limited, the relevant capabilities of, say, the human skeleton tracking algorithm (case A) or an intelligent language model training and inference capability (case B) could be implemented through native AI capabilities at the edge of network.

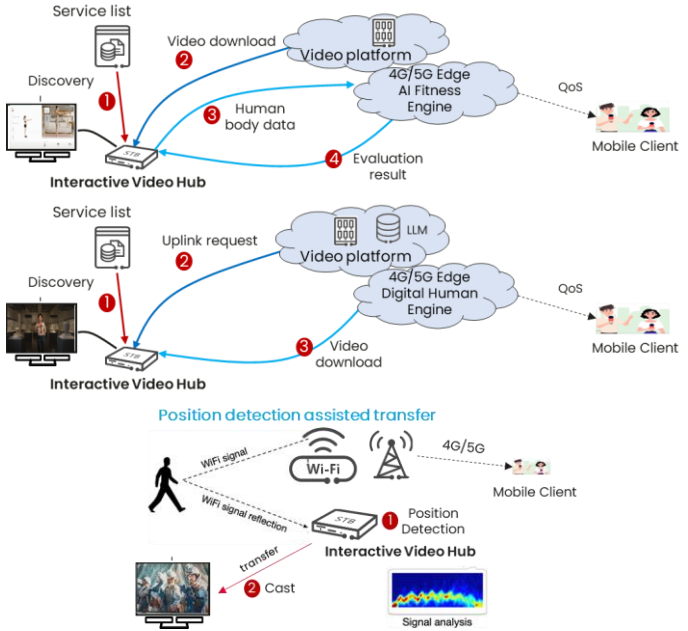


Figure 8. Possible deployment scenarios for case A (top), B (middle) and C (bottom)

For case C, users can seamlessly share media data and services between multiple different devices and seamlessly transfer data across their screens over different networks. Real-time sharing can be synchronized among multiple parties through transfer assisted by position detection.

High-level perspective on potential requirements

Connectivity

Ubiquitous massive access capability, with synchronous access for multiple parties, should be supported, both for users located indoors and outdoors across different domains of the network. High bandwidth for upstream and downstream access is required.

In general, low latency between users, and highly reliable QoS assurance are needed to ensure fast audience reactions. When user interaction is needed, content shared between them should be time synchronized, while minimizing the delay between them.

AI and computing capabilities

Because of limited capabilities on the device, the network should be able to provide computing capabilities to train and generate personalized models per user and implement 3D or immersive media processing and rendering capabilities, while sustaining acceptable end-to-end latency.

Transmission protocols

The transmission of multiple media should be interrelated and synchronized. Transmission protocols across multiple screens/users may be required.

Codecs for interactive media

Due to the fact that media data is not only video and audio, but also includes structured data for interaction, and all the data should be interrelated and synchronized, efficient codec formats for interactive media services may be required.

Multiview: Multiview video services

Multiview video services enable audiences to choose a different angle of view as their primary video by using a finger gesture on a touch screen device or by using the directional buttons on the remote control of an STB. To achieve this, multiple synchronized cameras are deployed in different positions to cover an event from different camera angles.

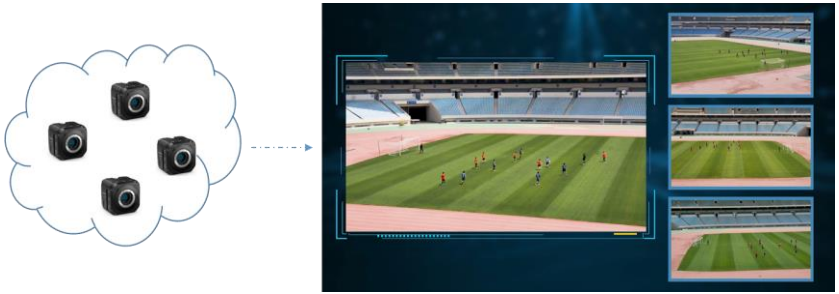


Figure 9. Overall end-to-end multiview video use case (within single device).

Two different experiences are possible in a multiview video scenario:

Multiview video within a single device

A user can watch the same programme from multiple angles at the same time and choose one viewing angle as the primary video on a single device. Synchronization is maintained for the viewer during the switch.

Figure 10. Multiview video within single device

Multiview video within multiple devices / companion screen

An viewer can watch the same programme from multiple angles at the same time on different devices, e.g. one device is a TV and the other is mobile phone. The audience may be able to have an experience of free viewpoint video on the mobile phone in this case. Synchronization among the devices is maintained for the viewer.

Figure 11. Multiview video within multiple devices

Architecture, relevant features and pre-conditions

Below are two examples of the production and distribution of multiview video.

Four synchronized cameras are deployed in different positions at a sporting venue, capturing the actions of the players from different angles. After passing through a media processing system, the streamed video is transported via the broadcast network and 5G network or via the 5G network with multicast/broadcast and unicast mode. TVs and phones can playback video from different angles synchronously.

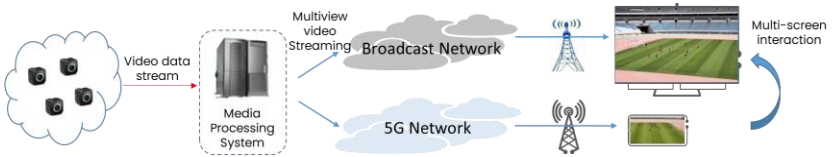


Figure 13. Production and distribution of multiview video via the broadcast network and 5G network

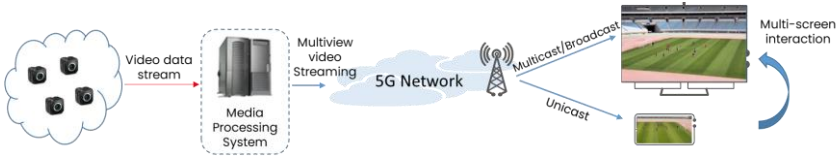


Figure 12. Production and distribution of multiview video via the 5G network with multicast/broadcast and unicast mode

High-level perspective on potential requirements

Users

Users should be able to watch a programme from different angles at the same time on a single device and choose one viewing angle as the primary video.

Users should be able to watch a programme from different angles at the same time on different devices and choose from among the available angles on each of the devices.

Codecs and protocols

Codecs and protocols should be selected to enable offering the content for different angles and to enable the synchronization among different views on a single device or multiple devices.

Required distribution QoS

Synchronization among multiview feeds should have a deviation of less than 120 ms.

Required QoE

Non-synchronization between the different views is not perceived.

Multiview: Stereoscopic 360° multi-camera

Audiences can view the action of scenes from various spots. This can be experienced with a single HMD or with a mobile device as a second screen to complement the main HD production transmitted for TV audiences.

An example is *Eufòria*, a singing talent show from the Catalan public broadcaster (3Cat), that aimed to find the best novice singer in Catalonia. In addition to the main HD TV production, four stereoscopic 360° cameras were set in different positions to allow users select the camera view. These views were live-streamed to personal computers, tablets, smartphones and HMDs.

Figure 13. Immersive 3D 360° multi-camera web application

5G devices, like smartphones or future HMDs, would leverage 5G technologies to optimize the user experience thanks to very low latency and a high bandwidth transmission, multicast distribution possibilities inherent to 5G as well as DRM (digital rights management) to protect the media content.

Architecture, relevant features and pre-conditions

A certain number of stereoscopic 360° cameras are set in different spots of a TV production studio or a venue (e.g. on the stage, among the jury, with the participants, in the audience or in the production control). The system allows the parallel encoding of all camera signals for streaming, perfectly synchronized, to computers, tablets, smartphones and HMDs.

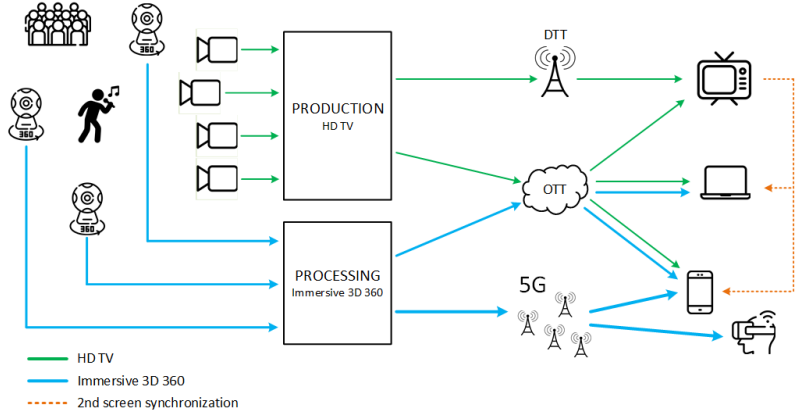


Figure 14. Production and distribution of Immersive 3D 360° multi-camera.

Some of the aspects that were not covered during this pilot were:

- Synchronization between broadcast delivery and the immersive streams.
- Quality of the immersive streams – low resolution and bit rate for this pilot
- DRM protection of the immersive content

High-level perspective on potential requirements

Users

Users should be able to access different immersive stereoscopic 360° views, choosing from among different viewing spots (multi-camera). From each spot, the user can choose the field of view (360°) moving their head or the device itself (HMD, smartphones, tablets) or using their fingers or mouse on fixed screens.

Codecs and protocols

Codecs and protocols should be selected to enable high resolution and high-quality immersive video, seamless switching of the immersive camera streams with audio continuity, and synchronization of the immersive content with the main HD programme, which could be delivered over broadcast networks (e.g. DVB-T/T2) or the internet. Protection of the content through DRM should be applied.

Required distribution QoS

The distribution stream would be produced using multi-quality profiles, and the maximum bandwidth required for the higher quality would be about 50 Mbps on the receiving device for a stereoscopic 360° 8K experience, while latency should ideally be low enough to achieve synchronization with the main TV broadcast signal.

Required QoE

Optimal image quality should be available for every device, especially to enable a stereoscopic experience on HMDs. There should be a sufficiently fast reaction when swiping a finger – scenes are changed seamlessly. No delay with the main HD TV broadcast is perceived by the user.

Virtual Reality (VR): VR Field of View (FOV)

VR FOV enables the provision of a high resolution (e.g. 8K) video experience to a user whose VR device is only decoding lower resolution content (e.g. 4K), therefore decreasing the required bit rates as well.

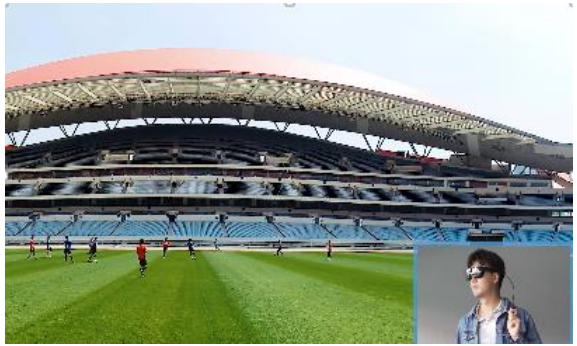


Figure 15. 8K VR FOV video experience

VR technology can be used in surgical guidance, live sports events, industrial design, shopping experiences, architectural models and interactive entertainment, to provide users with a more realistic visual experience of ultra-high-definition video. To provide high resolution, a wide viewing angle and rich details, 8K video would be preferred to 4K. However, there are two main challenges for 8K VR to be widely applied: (1) most VR devices can decode only 4K VR video; (2) typically, the bit rates for 8K VR video streams exceed 100 Mbps. FOV streaming is a popular solution to decrease the bit rates required to provide such services over existing networks.

Architecture, relevant features and pre-conditions

Deployment of 8K VR FOV can reuse existing CDN architectures and user devices.

Two encoding processes ensure the availability of lower definition VR panoramic content and higher definition 360° VR FOV content. The assets should be present at CDN nodes and served to users according to the selected view.

The device only decodes the FOV for the user's currently selected viewpoint, doing so instantaneously in high resolution, and therefore there is no need to transport the complete scene in that same resolution. When the user selects a new viewpoint, a new 8K resolution scene is shown in the FOV of the user.

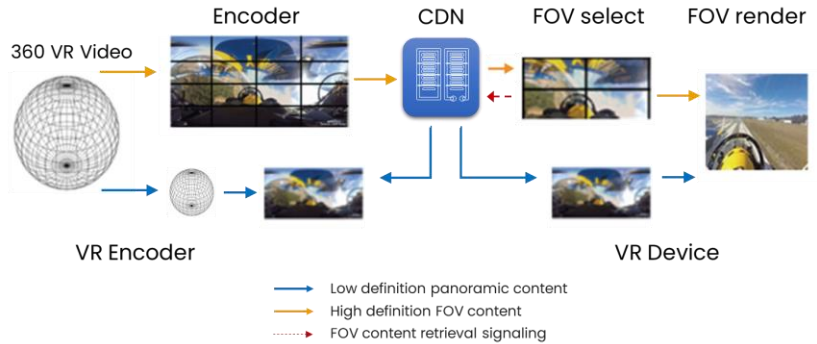


Figure 16. Content preparation and distribution of 8K VR FOV.

High-level perspective on potential requirements

Users

The user should be able to access immersive stereoscopic 360° views by moving their head in immersive devices (e.g. HMDs).

Codecs and protocols

Codecs and protocols should be selected to make it possible to offer content with the required different resolutions, i.e., high quality for FOV content and low quality for panoramic content, and seamless switching across the FOV content for different viewpoints.

Required distribution QoS

Compared with standard 8K VR, using 8K VR FOV can save 50% to 60% bandwidth, i.e., the maximum bandwidth required for the higher quality would come to about 50 Mbps for a typical 8K VR video.

Maximum action latency is less than 50 ms during the change of viewpoint.

Required QoE

No delay must be perceived during the change of the viewpoint.

3D Scene-based media: Glasses-free 3D video

3D Scene-based media are motion pictures made to give an illusion of three-dimensional solidity, usually with the help of special glasses worn by viewers or on a 3D display that does not require the use of such glasses.

Figure 17. Visualization of a glasses-free 3D video experience

There are two major glasses-free 3D video display experiences:

1) Single-view 3D display experience. This usually applies to a mobile phone or tablet and allows a single user to watch a 3D video from the appropriate viewpoint. With an eye-tracking system, the user can continue to watch the 3D video when changing the viewpoint, within an applicable range.

2) Multiview 3D display experience: This usually applies to a TV set or outdoor billboard and allows multiple users watch a 3D video from multiple viewpoints simultaneously.

Architecture, relevant features and pre-conditions

Glasses-free 3D is any method of displaying stereoscopic images (adding binocular perception of 3D depth) without the use of special devices.

There are various Glasses-free 3D display technologies, e.g. parallax barrier, lenticular lens, light field displays, etc., to offer better image quality with high resolution. Eye tracking and multiple views are two of the approaches that have been utilized to accommodate the motion of the viewer.

The main elements of the architecture are the capture cameras (from single setups to arrays) and encoders. A media-processing subsystem is responsible for the necessary conversion and synthesis of the different 3D views. A video platform is responsible for receiving indications of the user position (in case of a single view) and delivering the appropriate viewpoint.

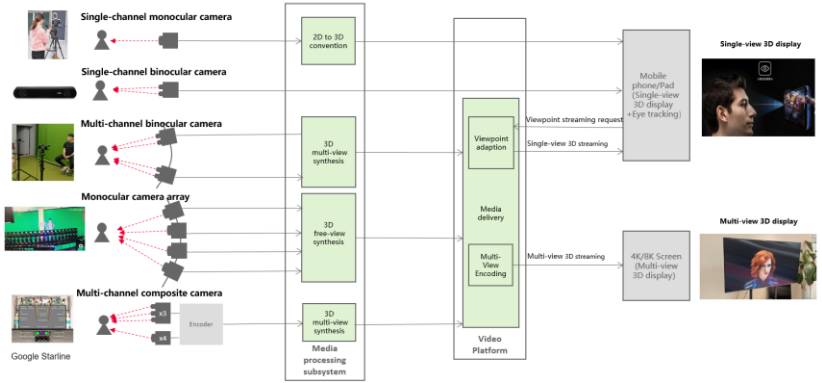


Figure 18. Glasses-free 3D video production and display

High-level perspective on potential requirements

Users

Users should be able to watch 3D video without the help of special glasses.

Codecs and protocols

Content delivery and streaming protocols should support high resolution, 3D multiview/free-view synthesis, eye tracking and viewpoint streaming retrieval.

Required distribution QoS

Single-view 3D display requires at least twice the bandwidth of 2D video. Action latency should be less than 50 ms if retrieval of the viewpoint stream is required.

Multiview 3D display requires three times or more the bandwidth of 2D video.

Required QoE

No delay must be perceived when changing the viewpoint within an applicable range.

Volumetric Video: XR for Media Production

The starting point is a TV production consisting of multiple participants physically present in a studio and a set of remote participants, e.g. debating or being interviewed. Today, the communication between the local and the remote participants is done through audio/video conferencing. On TV, remote participants are either not visible, or inserted as picture-in-picture.

By leveraging volumetric video acquisition, coding, transport, and restitution in AR glasses, local and remote participants can see each other as if they were all locally present. This may be achieved by a seamless merging of the volumetric representations of both remote and local participants. The final TV service may consist of both a legacy 2D linear TV service with remote participants seamlessly merged within the studio, but also an XR 6DoF experience for viewers equipped with XR headsets, therefore enhancing the quality of experience of viewers and participants.

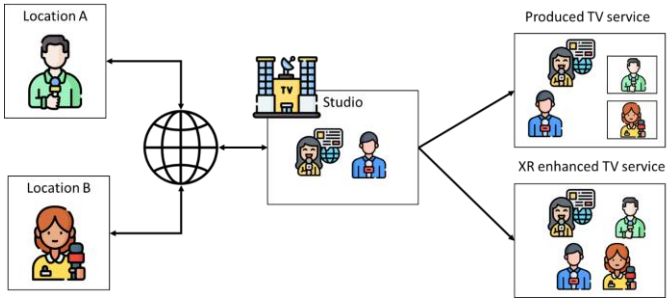


Figure 19. TV production with picture-in-picture (top) and enhanced XR service (bottom)

Computer graphics are widely used today to generate overlays or to embed additional content. This is a rather static, basic approach and is not tailored to the user experience in different devices.

Communication between local and remote participants in, for example, an interview is limited by the 2D capture of the remote participant inserted in the video service and the return channel, which suffers from a noticeable delay.

It is proposed to leverage XR technologies to increase the immersiveness of the production by introducing the following features in the workflow:

1. Live volumetric acquisition to capture remote participants and the studio.
2. Augmented Reality glasses: to embed remote captured participants in journalists' fields of view and enhance communication between them.
3. XR-capable production: to embed remotely captured volumetric elements in a 2D flow, in a seamless manner.
4. Live volumetric video coding: to enable delivery of XR video services from the remote location to the studio and distribution to the end user.

Adopting such technologies requires additional processing, technological bricks and connectivity that can be provided by the 5G system.

Architecture, relevant features and pre-conditions

The volumetric acquisition system enables capture both in the studio and remotely. Acquired data is transmitted to an edge server responsible for processing and building a 3D model of participants and studio. The constructed 3D models are transmitted to the split-rendering engines, and the production control room through wired infrastructure. Legacy cameras and microphones reach the control room via wired infrastructure.

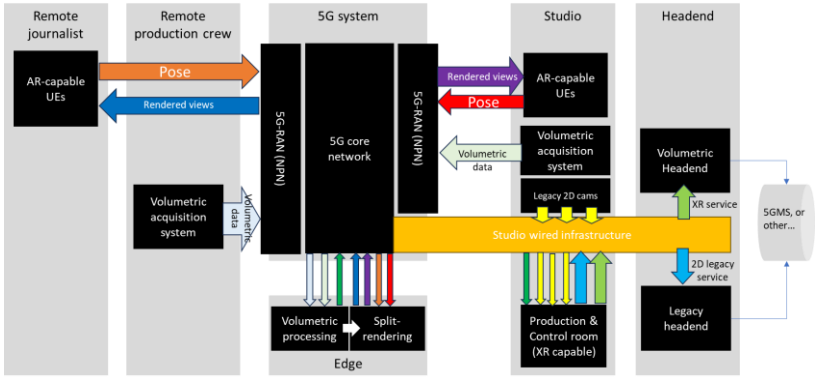


Figure 20. High level XR-enhanced production setup

All AR-capable devices require split-rendering, needed to reach the typical light, small and regular-glasses form factor. The pose information is continuously captured and sent to the split-rendering engines, which return rendered views.

Wireless connectivity for capture and for device communication is achieved through 5G, able to meet throughput and latency requirements.

The control room mixes XR and 2D sources, generating two services:

1. A legacy 2D service, to be viewed on regular devices (TV, smartphone, …)
2. An XR service, to be viewed on XR devices (AR glasses, VR headset, …)

The delivery of these services is out of scope, and may be done with 5G Media Streaming, or any suitable existing delivery protocols or paths.

High-level perspective on potential requirements

Remote participants are captured with multi-camera and depth-acquisition systems and smartphones. The stage is equipped with legacy cameras.

Participants are equipped with in-ear devices and microphones, for real-time audio communications, and with AR glasses. The form-factor shall be that of regular glasses to maintain the viewing experience for the end viewer.

The AR glasses can offload the intensive rendering task to an edge cloud.

Communication networks are available, including 5G non-public (NPN) or public networks, or a local WI-FI network, or a combination of both. The required multimedia processing can be achieved on today's platforms, typically on server equipped with GPUs. Legacy video codecs can be used in combination with e.g MPEG Visual Volumetric Video-based Coding (V3C) and related metadata to provide all-in-one data streams. Communications with low enough delay can be achieved with RTP.

The following requirements need to be supported to enable the service:

1. The communication between the participants, and their interactions should be seamless, with almost unnoticeable latency. This includes video, audio and volumetric bi-directional data flows.
2. The compute and network infrastructure should support required multimedia processing with minimal latency introduced.
3. Quality of acquired volumetric data, including audio and video (resolution, accuracy, synchronization) should be high.
4. Insertion of volumetric assets in the production workflow should be spatially and temporally seamless.

Free-Viewpoint Video: Live, Time-freeze, space-shift

In conventional video, fixed and cable-suspended cameras capture the action. Audiences can only watch discrete images from limited viewpoints chosen by the service provider during the changing of the viewpoints. With free-viewpoint video, audiences can be provided with an immersive media experience where the action can be experienced with continuous scenes from various angles during the changing of the viewpoints. Free-viewpoint video is typically used for sports events, concerts, product launches, films, etc.

As an example, in a live tennis game, a set of synchronized cameras are deployed on one side of the court. The cameras capture the action from different angles. All images captured by the cameras are processed into a single unified piece of content. For example, video frames from adjacent angles with the same timestamp will be stitched. And for some virtual viewpoints, processing of frame interpolation could also be needed. An audience in the stadium could actively chose any angle to view the action of the scenes by using a finger gesture on a touchscreen device. A remote audience at home could chose the viewing angle by using the left and right directional buttons on the remote control of an STB.

Figure 21. Capture, viewing on phone and on TV

Free-viewpoint live streaming: the user can select any viewpoint to watch from the live streaming video. The user can rotate their viewing position to a different angle at any time, with a one finger gesture on a touchscreen device or using the left and right directional buttons on the remote control of an STB.

Time-freezing highlights: the user is provided with a highlights video consisting of a time-frozen scene, allowing the user to view the stopped scene from various angles.

Space-shifting highlights: the user is provided with a highlights video consisting of various angles of a spatial video captured over a period of time. The service provider may insert this content while the normal video is playing or prompt the user to download the highlight clip.

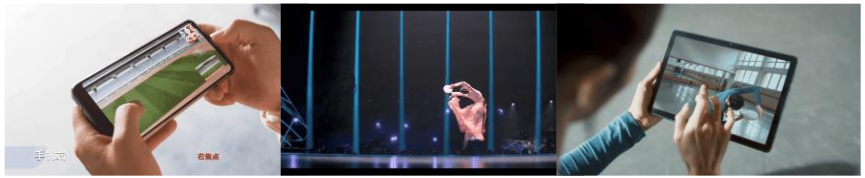


Figure 22. Free-viewpoint live streaming; Time-freezing highlights; Space-shifting highlights. Click the following link to see the animated graphics: https://www.5g-mag.com/post/media-services-beyond-2d-use-cases

Architecture, relevant features and pre-conditions

An array of multiple (e.g. tens of) synchronized cameras is deployed in a linear, circle, ellipse or butterfly-shaped rig to record the main actors in a 360° view, presenting a video that allows smooth 3D rotation, and can generate a highlight-moment video. In order to enable the generation of a dynamic video scene with any new viewpoint chosen by the user, all the stream data from the array of cameras shall be processed, transported and stored, and the live streaming of the new viewpoint shall be provided to the user.

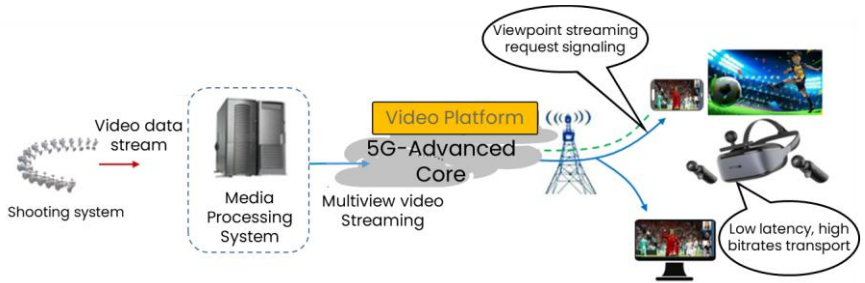


Figure 23. Production and distribution of Free-viewpoint Video

High-level perspective on potential requirements

Required production QoS

Bit rates for the processed media streaming can be up to 3 Gbps.

Required distribution QoS

Bit rates and latencies should be sufficient to render the viewport within the immersive limits. The bit rate can be up to 30 Mbps and latency lower than 50 ms.

Required QoE:

There must be a sufficiently fast reaction to swiping of the finger to ensure scenes are changed seamlessly.

Requirements on distribution system:

Efficiency: the video content should be distributed in an efficient way, i.e. the bandwidth should be used in an efficient way as much as possible, especially in the case when there is a large audience, when group-cast-like distribution may be considered.

Low latency: the distribution latency is critical to the user experience. One way to achieve acceptable latency is by placing the distribution node close to the audience, for example, distributing the video content based on edge computing architecture.

Scalability: in order to distribute video content in a scalable way, some relay nodes could be introduced between video producer and audiences to achieve scalability, for example third-party networks, independent of the video producer and audience, can be leveraged to relay the content.

Social: Watch Together

Social TV provides an experience of social video by integrating real-time communication systems with an existing TV system (e.g. a DVB system). Participants of the social TV experience can have a video chat about the TV programme that they are watching simultaneously.

There are several OTT apps supporting this kind of experience, e.g. BT Sport's Watch Together and TikTok's VR Live. Here we describe a social TV use case that involves integrating an existing TV service network and operator network.

Figure 24. Concept of a Watch Together service

All the participants of the social TV experience are subscribers to a TV service provider. One participant can share the TV programme (linear or on-demand) with other participants and they can chat with each other when they are watching the programme as if they were sitting in front of the same TV. Synchronized playout is key to these experiences.

Architecture, relevant features and pre-conditions

In the architecture show below, a mobile phone or STB supporting both a TV video service and IMS (IP Multimedia Subsystem) can provide the social TV experience to the user with the support of the traditional TV video platform and IMS combination. IMS is a standards-based architectural framework defined by 3GPP for delivering multimedia communications services such as voice, video, and text messaging over IP networks. Recently, the data channel technology defined for WebRTC has been introduced into IMS to transfer the data of data channel applications between end-user devices or between devices and the network.

The devices (e.g. mobile phone or TV) have the capabilities to access the same TV video service via the 5G network. One of the devices acting as the chair initiates an IMS call or IMS conference call including a data channel created with the other participants.

The device acting as chair shares the screen with the other participants during the preview of TV programme information via the data channel.

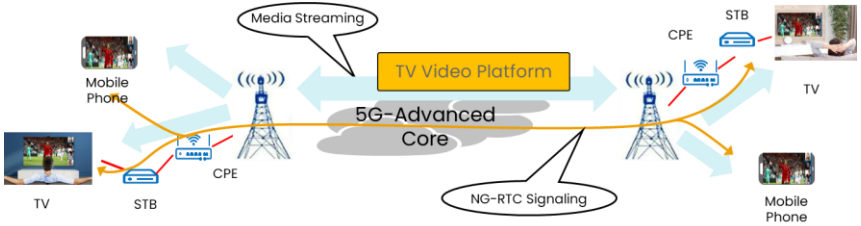


Figure 25. Social TV architecture

Once a programme is selected, the device acting as the chair shares the location of the programme (e.g. URL) with other participants via the data channel and all the participants access the selected programme based on the shared location.

The device acting as chair controls the playout (e.g., play, pause, synchronization, etc.) by sending the signalling via the data channel.

High-level perspective on potential requirements

Users

Users should be able to have a video chat about the same TV programme, which they are watching simultaneously on a mobile phone or TV.

Codecs and protocols

Protocols support the establishment of a communication session among the participants and the sharing of the individual screens of the participants and the main video Protocols support to control the playout of the video by one participant.

Required distribution QoS

Audience Drift Gap is less than 100 ms. End-to-end latency between the users is less than 500 ms.

Required QoE

Participants perceive their viewing of the TV programme as in sync with each other.

Definitions

Degrees of Freedom (Quoted from 3GPP RT 26.928)

A user acts in and interacts with extended realities as shown in the following figure. Actions and interactions involve movements, gestures, body reactions. Thereby, the Degrees of Freedom (DoF) describe the number of independent parameters used to define movement of a viewport in the 3D space.

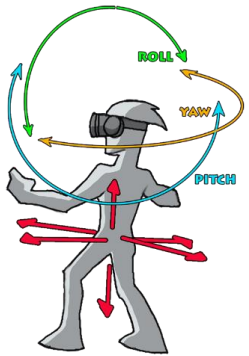


Figure 26: Different degrees of freedom for a user in extended realities

3DoF: Three rotational and un-limited movements around the X, Y and Z axes (respectively pitch, yaw and roll). A typical use case is a user sitting in a chair looking at 3D 360 VR content on an HMD.

3DoF+: 3DoF with additional limited translational movements (typically, head movements) along X, Y and Z axes. A typical use case is a user sitting in a chair looking at 3D 360 VR content on an HMD with the capability to slightly move his head up/down, left/right and forward/backward.

6DoF: 3DoF with full translational movements along X, Y and Z axes. Beyond the 3DoF experience, it adds (i) moving up and down (elevating/heaving); (ii) moving left and right (strafing/swaying); and (iii) moving forward and backward (walking/surging). A typical use case is a user freely walking through 3D 360 VR content (physically or via dedicated user input means) displayed on an HMD.

Constrained 6DoF: 6DoF with constrained translational movements along X, Y and Z axes (typically, a couple of steps walking distance). A typical use case is a user freely walking through VR content (physically or via dedicated user input means) displayed on an HMD but within a constrained walking area.

Audience Drift Gap :

Time difference between the first user to see a frame of media and the last user to see that same frame of media.

Summary

This report summarizes a collection of existing and new use cases and services beyond 2D video. It also provides an overview of architectures, features and high-level requirements for those use cases.

The use cases were classified under seven service types: interactive video services, multiview services, virtual reality (VR) experiences, 3D scene-based media, volumetric video, free-viewpoint video and social video. The table below shows the nine use cases presented, indicating which of six key characteristics apply to each.

	Fidelity	Interactivity	Immersiveness	Sociability	Intelligence	Ubiquity
Massive Interactive Live Events	X	X	X	X		X
Interactive Hub	X	X	X	X	X	X
Multiview Video	X	X				X
Stereoscopic 360° Multi-camera	X	X	X			X
8K VR FOV	X	X	X			
Glasses-free 3D Video	X	X	X			
XR for Media production	X	X	X	X		X
Free-viewpoint Video	X	X			X	
Watch Together	X	X	X	X		X

Note that the extent to which the referred use cases are supported or not by current technologies, in particular 3GPP and media related specifications, is not addressed in this document and may be part of future work. A consolidated list of detailed requirements is also left for future work.

www.5g-mag.com

This is a report produced by the 5G-MAG Workgroup UC (Use Cases, Requirements and Opportunities).

Version of the report: v.1.0

Date of publication: May 2024

This 5G-MAG Report can be downloaded from www.5g-mag.com/reports

Feedback from the industry is welcome through https://github.com/5G-MAG/Requests-for-Feedback

Published by 5G-MAG | May 2024

5G-MAG Media Action Group Association
17A L'Ancienne-Route
1218 Grand-Saconnex (Switzerland)

info@5g-mag.com • www.5g-mag.com