120th MPEG Macau SAR, China, 23 - 27 October 2017, Meeting Report

Panos Kudumakis
qMedia, Queen Mary University of London

Contents

1 69th Engineering Emmy Awards Recipients Announced

- **The Charles F. Jenkins Lifetime Achievement Award**
  Honors a living individual whose ongoing contributions have significantly affected the state of television technology and engineering.

  Recipient: **Leonardo Chiariglione**

  As founder and chairman of Motion Picture Experts Group (MPEG), Leonardo Chiariglione has led MPEG in setting the worldwide standards for digital video compression and transmission. He will be honored for his pioneering technology and innovation efforts in the field of video compression.

- **Engineering Emmys**
  Presented to an individual, company or organization for engineering developments that considerably improve existing methods or innovations that materially affect the transmission, recording or reception of television.

  Recipient: **High Efficiency Video Coding**

  The development of High Efficiency Video Coding (HEVC) has enabled efficient delivery in ultra-high-definition (UHD) content over multiple distribution channels. This new compression coding has been adopted, or selected for adoption, by all UHD television distribution channels, including terrestrial, satellite, cable, fiber and wireless, as well as all UHD viewing devices, including traditional televisions, tablets and mobile phones. The Emmy goes to the Joint Collaborative Team on Video Coding (JCT-VC), a group of engineers from the Video Coding Experts Group (VCEG) of the International Telecommunication Union (ITU) and the Moving Picture Experts Group (MPEG) of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) for the development of High Efficiency Video Coding.

Related Articles
1 Television Academy Announces Recipients of the 69th Engineering Emmy Awards
2 IEC, ISO and ITU receive Primetime Emmy award for excellence in video engineering
3 Leonardo Chiariglione, On my Charles F. Jenkins Lifetime Achievement Award

2 Network-Based Media Processing

MPEG has developed various technologies for multimedia coding and transport, such as AVC/HEVC, 3D audio, MPEG-2 TS, ISOBMFF, DASH and MMT. These technologies have been widely adopted and are heavily used by various industries in various applications, such as digital broadcasting, audio and video streaming over the Internet, in mobile terminals, etc.

In order to develop standardized and efficient solutions for network-based media processing (NBMP), especially given the recent increase in demand for distribution of MPEG media in next generation network environments

such as 5G, MPEG evaluates and addresses the current limitations of available standards in the MPEG media distribution area including taking considerations of processing units in networks and challenges in emerging network environments into account.

NBMP is a framework that allows service providers and end users to describe media processing operations that are to be performed by the network. NBMP describes the composition of network-based media processing services out of a set of network-based media processing functions and makes these network-based media processing services accessible through Application Programming Interfaces (APIs).

An NBMP media processing entity performs media processing tasks on the input media data and the related metadata. NBMP also provides Control Functions that are used to compose and configure the media processing. In addition, NBMP provides transport methods for communication between media source and media processing entities.

In particular, the NBMP framework will define interfaces between the Media Source and the Media Processing Entity and among the Media Processing Entities, that will allow users to access the framework, configure the media processing entity, upload/stream media data to the network for media processing, and utilize the processed media and the resulting metadata in real-time or in a deferred way.

The media and metadata formats that are used between Media Processing Entities in a media processing pipeline are also within the scope of the standard as well as the Workflow description that is used to orchestrate media processing entities and to compose media processing services into a pipeline of media processing entities.

The following diagram depicts the NBMP architecture that will be used as a reference architecture to scope the NBMP work.

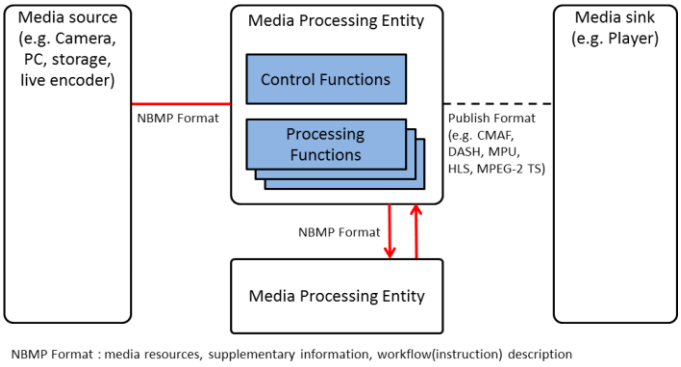

Figure 1 - Potential framework for Network-Based Media Processing system.

The NBMP framework consists of two generic functions inside media processing entities: Processing Functions, and Control Functions.

- Processing Functions: provide functionalities for media processing and analysis given control instructions from the control functions. These functions are considered to generate processing output information, which include the following functions:
    - Media processing function: the core media processing function which performs processing of the input media that can generate output media or metadata. Examples of media processing are follows; content encoding, decoding, content encryption, content conversion to HDR, content trans-multiplexing of the container format, streaming manifest generation, frame-rate or aspect ratio conversion and content stitching, etc.
    - Analytics function: provides functionalities for analysis of the logged information and makes the analysis report available via request.

- Control Functions: provide functionalities for controlling and management of the media processing tasks and workflows, and how input media is processed into output media published into the media sink. A key control function is the “service/workflow manager” function.
    - Service/Workflow function: provides functionality for composition of media processing workflows by chaining a set of media tasks. This includes matching the output of each media processing entity to the input of the succeeding media processing entity.
    - Monitoring Function: provides functionalities for monitoring the processing pipeline, and to guarantee media processing task execution, correctness, or to detect failures during media processing.
    - Logging Function: logs information about media processing and/or services

- Pub/Sub Function: provides functionality for messaging and information exchange, including content retrieval and publishing. This may be used to trigger processing to start at one media processing entity after processing ends in a previous media processing entity.
- Security functions: provides functionalities for ensuring user's and content security, including user authentication, content encryption, and other functionalities, as necessary

Use cases

- UC#1: Network assistant VR stitching
- UC#2: Live Media Ingestion
- UC#3: Intelligent video up-scaling
- UC#4: Augmented Video streaming
- UC#5: Mobile Edge Encoding for the adaptive streaming
- UC#6: Intelligent user centric broadcasting
- UC#7: Interactive media services
- UC#8: Media processing for Vehicles
- UC#9: Media aware caching

Output Documents
N17262 - Use cases and draft requirement for NBMP (v2)
N17263 - Draft Call for Proposals on Network-Based Media Processing

3 MPEG-I Audio Use Cases

In Virtual Reality use cases an audio-visual scene is rendered to a user's head-mounted display and headphones. However, as the user's head turns or the user's position changes, the scene is rendered consistent with the virtual reality. For audio rendering, this involves changing the orientation of the sound stage as the head orientation and user position changes and perhaps also altering the sonic rendering, e.g. with varying degrees of reverberation, as a consenquence of changes in user location with a space. Since the rendering is of a virtual world, all sound objects and sound processing can be known prior in advance.

Augmented Reality use cases are similar in presentation, but since they involve a virtual world (e.g. one or more virtual sound objects) imposed on true reality, aspects of the real world may not be know in advance. For example, the user moves into a physical space with new and different sound properties. Such sound properties (e.g. room impulse response or reverberation) may have to be inferred from other sensors, e.g. position or visual.

Use cases

1. **Multi-viewpoint 3DoF with monoscopic/stereoscopic 360 video and 360 audio content**
   The viewer is presented with a multi-viewpoint 3DoF experience. The viewer is given the opportunity to change his/her 360 degree viewpoint, selected from amongst a set of multiple predefined (fixed) viewpoints within the content. Changing between the fixed viewpoints can be either controlled manually by the viewer (using some kind of local controller), or automatically through specified metadata (e.g. a director's cut or guided viewing). The corresponding audio for the current viewpoint location will also be delivered and rendered accordingly.

   Required features:
   Multi-viewpoint

2. **Single viewpoint 3DoF+**
   Single viewpoint 3DoF+ with full 3D 360 video is an enhanced experience of the fixed single viewpoint use case. This use case provides a realistic, natural full 3D 360 audio/visual experience, where the rendered content provides a natural 3D representation depending on all head rotation orientations (yaw, pitch, roll) of the viewer. As a 3DoF+ experience, this single viewpoint content also provides a limited amount of motion parallax, enough to give the viewer a sense of natural depth whilst changing his/her view within the scene, as well as the capability for small head translational movements. Small head translational movements are defined as movements which can be achieved whilst the viewer is in a seated position, without the use of the lower body. In this manner, this content is comparable to a restricted 6DoF experience, where translational movements are limited to small head movements. Audio is also rendered accordingly based on the user's head position and orientation, including possible sonic occlusion, if perceptually relevant.

   Required features:
   3DoF+ navigation with associated interface to renderer
   Audio rendering responsive to small head movements

3. Single user in VR environment, 6DoF

A single user can navigate within a VR environment. It may be that there is no or simple interaction, i.e., interaction that has been predetermined by the content creator. Interaction may be responsive to user coordinates, orientation, or user/object proximity, e.g., if I am near an object and look at it, it begins to "talk".

User interactions with VR objects are rendered in real-time. It may be that only audio components are presented in VR environment, e.g., AR augmented with audio (e.g., without augmented visuals) may be used for AR audio advertisement.

Required features:
6DoF navigation with associated interface to renderer
Audio rendering responsive to user movements
Navigation within "background" audio
Navigation within and around VR audio objects
Interactions with VR objects

4. Multiple users in VR environment, 6 DoF

Several users can navigate simultaneously within a shared VR environment. User interactions with VR objects and between VR users are shared in real-time among all users.

Required features:
Interactions with VR objects
Rendering of other users in the virtual environment, including possible speech or audio from other users.
Interaction between users, including voice communications. Voice communications between users must be low-latency.
Synchronization of interactions among users, e.g. lip movement and lip synch of speech.
Synchronization of audio and video of users and the scene.

The following table shows that use-cases are matched with the proposed features.

Features under consideration	**Multi-viewpoint 3DoF with monoscopic / stereoscopic 360 video contents**	**Single viewpoint 3DoF+**	**Single user in VR environment, 6 DoF**	**Multiple users in VR environment, 6 DoF**
Binaural rendering over headphone	O	O	O	O
Low Latency for motion-to-sound interaction	O	O	O	O
Low Latency for communication				O
Accurate 3D localization	O	O	O	O
Occlusion by objects		maybe	O	O
Raditation pattern of a sound source		maybe	O	O
Doppler effect			O	O
Modeling of acoustic space			O	O
Soundfield interpolation/extrapolation			O	O

Output Documents
N17176 - MPEG-I Audio Use Cases
N17177 - MPEG-I Audio Architecture and Evaluation Procedures for 6 DoF
N17253 - Thoughts on Test Material and Encoder Input Format
N17254 - Workplan for MPEG-I 6 DoF Audio

4 MPEG-I Audio Architecture and Evaluation for 6DoF

It is understood that 23090-4, MPEG-I Audio will be based upon 23008-3:201x, MPEG-H 3D Audio Second Edition.

MPEG-H 3DA includes all coding technology that is necessary to carry 3D Audio content for 6DoF by supporting channels, objects and HOA. The core waveform carriage and essential metadata part of MPEG-H 3DA will be employed also by MPEG-I 6DoF. The "back-end" part of MPEG-H 3DA is dedicated to rendering these waveforms to any flexible output speaker configuration or binaurally to headphones. MPEG-H 3DA supports rotation of the presented sound scene in response to user interaction as indicated by (yaw, pitch, roll), or 3DoF. However, it is desired to support VR/AR applications that have full 6DoF of user interaction. To achieve this, additional user interaction as indicated by (x, y, z) translation has to be added to MPEG-H 3DA.

The VR and AR use-cases emphasize the importance of headphone rendering. In addition to headphone-based listening, a key component of the VR and AR 6DoF use-cases is the incorporation of *time-varying interactive rendering of linear translation with sufficiently low motion-to-rendered-output latency* compared to that currently supported by MPEG-H 3DA.

Thus, the new additional technology to be defined for MPEG-I 6DoF Audio consists of:

- New interactive rendering technology, also supporting linear translation, for objects (including channel, i.e. objects with a fixed placement) and HOA which fully supports the VR/6DoF use cases.
- Additional metadata needed to support the VR/6DoF applications beyond what is already available from MPEG-H 3D Audio. This can be packaged in an extra bitstream container such that the content is still compatible to MPEG-H 3D Audio rendering.

In view of this, evaluation of newly proposed technology consists of an evaluation of the decoding & interactive rendering process where each technology candidate would essentially perform:

- decoding of a common reference MPEG-H bitstream representing the content elements as any combination of channels, objects or HOA
- decoding of newly defined individually generated VR/6DoF extension bitstreams containing the additional metadata that is needed for VR/6DoF application
- newly defined interactive rendering of the decoded item to the output devices (headphones and/or loudspeakers)

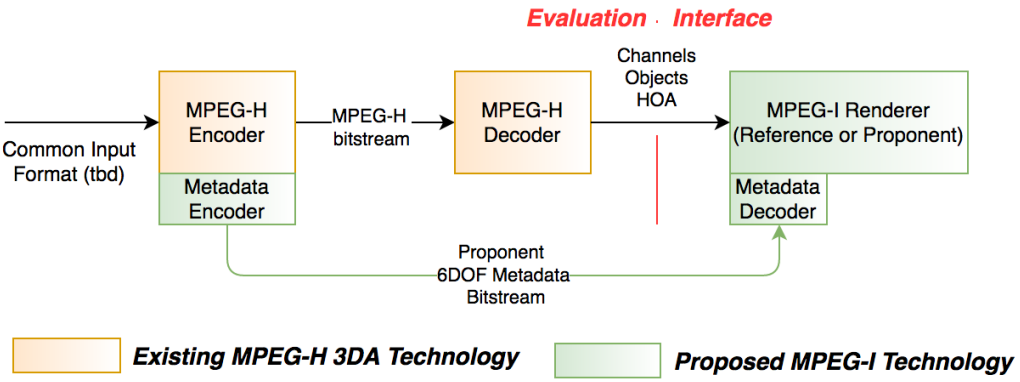

Figure 2 - MPEG-I Audio Architecture for 6DoF.

Interaction and Presentation

The following subsections discuss issues of the degree of interactivity with the environment that is provided by the test procedure. This alludes to whether the user should be free to navigate naturally and unhindered through the audio scene with 6DoF or the degree to which their translational and orientation movements should be prescribed as part of the evaluation set-up.

- The fully interactive navigation option may be more restricted in terms of area and complexity of audio scene due to physical constraints (e.g. room size the user is in, tethering cables etc).
- Degree of repeatability over subject movement (desire: large repeatability). This may be more achievable with a predefined translational track which clearly compromises freedom of movement and realism / immersion even if the user still has 3DoF+.
- Degree of realism including body/head motion cues (desire: large/maximum realism / immersion). This may lead to increased variance of user opinions due to a lack of control over what users actually listen to and in what order.

1. **Full Subject Motion with and A/V Presentation**

The overall goal of a VR/6DoF technology is to create a ***convincing illusion*** of "the user being there" with a feeling of total ***immersion (presence)***. This feeling of immersion is generated by the user's brain if and only if all of the relevant cues reaching it are sufficiently consistent and consistent with the experience of daily reality. This includes the perception of body (and head) self-motion and the consistent low-latency reaction of the A/V presentation. In other words, it requires the action of body/head movement and reaction by the A/V presentation (see M41103).

In addition to audio cues, the following cues contribute to the feeling of immersion:

- The influence of ***head & body movement*** are considerable (see the low quality of illusion with untracked / tracked binaural audio and / or video). Similarly, the physical use of hand-held controllers that may be used by the user to position virtual objects within the virtual space by dragging them around is very effective in creating a convincing illusion. In this case, it is the perception of physical position & motion of the arm/hand together with an appropriate rendered A/V response that are powerful supporting cues to the brain
- The influence of the ***visual presentation*** is considerable (and maybe even more powerful that the audio-only localization cues).
- Attempts to test with reduced body movement will significantly reduce the feeling of user presence or immersion (3D Audio cues presented alone are weak ones, movement cues are strong and visual cues are strong, too).
- Conversely, it is quite possible that visual cues may be so strong that they mask audio defects and artefacts that would be otherwise apparent in an audio only presentation or application (e.g. The McGurk Effect).

It is thus envisioned that ideally users should be permitted to exercise full and unhindered body motion (i.e. head movement, body movement and (x,y,z) translation) when systems under test are presented for subjective evaluation. Obviously, this requires real-time decoding and rendering systems. Considering these facts, it is recognized envisioned that testing with visual presentation and full and unhindered user body movement is desirable in the interests of immersion and VR realism.

2. **Partial Subject Motion and A/V Presentation**

It is, however, also recognized that certain 6DoF application scenarios may not include full body motion (e.g. sitting in front of a computer and navigating via joystick), and that rendering for a predefined user path may provide more reproducibility. It is also recognized that presentation of a matching visual component will require extra effort.

As an alternative to the user being able to move in 6DoF, the audio soundscape could move consistently around the subject (see M41116). The subject is then permitted to move in 3DoF+ with respect to that moving frame of reference to judge the quality of experience.

It is proposed that the moving soundscape is implemented at the decoder/renderer side by a set of composite coordinates formed by summing a set of pre-determined movements (x, y, z and yaw, pitch, roll) synchronized with the source signal and the 3DoF+ movements under the control of the subject. Clearly the pre-determined translational movements will be designed to position the subject close to key active sound sources within the audio soundscape.

This approach will require significant care if applied to joint audio and video presentation to prevent sensory conflicts and resultant VR sickness. In joint audio and video testing however, the visual cues or instructions necessary to ensure that the subject is viewing the relevant parts of the visual scene will probably mean that this approach is probably less useful.

3. **No Subject Motion and A/V Presentation**

Yet another alternative to the user being able to move in 6DoF is for all user translation and orientation to be pre-determined (see M41688). It should be investigated whether this method could

- improve the reproducibility aspect
- improve the evaluation "coverage"
- overcome potential limitations imposed a need for real-time rendering. This may also lead to increased involvement of companies in MPEG-I audio development
- cover the use-case of "the 1st person view mode of a viewport controlled by the 3rd person"

The user path taken in this evaluation mode could be "recorded" from actual subject motion in the full subject motion method, or alternatively the motion data could be simulated.

In addition to fully interactive A/V testing, it has also been recognized (M41783) that there are relevant audio-only AR use cases with natural user sight and a headphone-based audio presentation that is reactive to user position and orientation. For such scenarios, and if isolated evaluation of audio cues is desired, audio-

only evaluation is envisaged as an alternative way of testing. Furthermore, other methods for testing with predefined user paths have been proposed (M41116, M41688) and are subject to further investigation.

Common Evaluation Platform for MPEG-I 6DoF Audio

Based on contributions M41709 and M41711, a common platform for development and evaluation is explored and further developed in a joint effort within the Audio Subgroup. The platform uses commonly available hardware (PC) and software components (Windows7, Unity, MAX/MSP7, UniOSC, …) and supports:

- Low latency user motion tracking and HMD visual display
- Presentation/rendering of MPEG-I Audio VR 6DoF content, including
    - Switching between different VR Audio renderers
    - Presentation/rendering of associated visual content
    - Collecting users' gradings by a test GUI (MUSHRA-VR)

The following diagram shows a top-level overview of the current system implementation. The workflow may be broken down into three modules: VR hardware, visual rendering engine, and audio rendering engine. Graphical and Audio components are currently rendered independently but are synced via OSC (Open Sound Control) data connection.

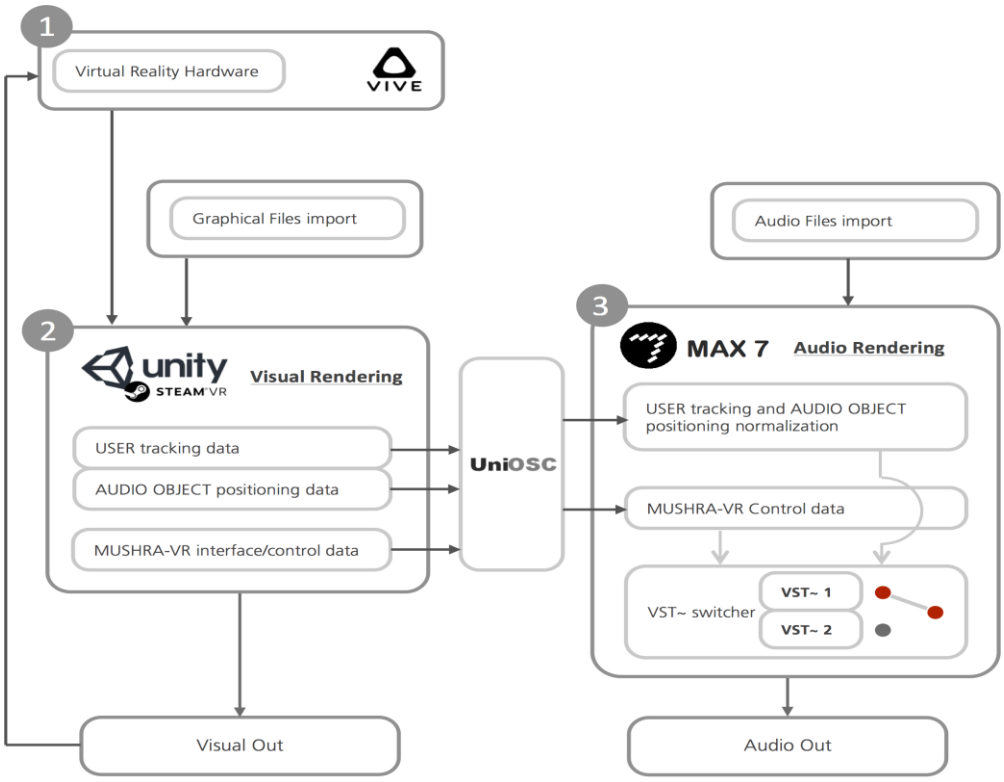

Figure 3 - Common Evaluation Platform for MPEG-I 6DoF Audio.

Subjective Evaluation Methodology

Based on the experience with MPEG-H 3DA, it has been agreed that a comprehensive and realistic evaluation is best performed using a single figure of merit (similar to the MUSHRA basic audio quality as in the case of MPEG-H 3DA) and that evaluating different aspects of the subjective quality will be performed by choosing an appropriate test set that contains test items which stress specific critical aspects of rendering.

It is furthermore proposed that this subjective figure of merit is based on the concept of *Quality of Experience* (i.e. QoE points) together with appropriate testing instructions (defining the application scenario to the user) and test set.

It is furthermore suggested to re-use the well-proven paradigm of *near-instantaneous switching* as it has been used successfully for a long time within MUSHRA to switch in real-time between different VR renderings – in other words to employ a "MUSHRA-VR" test methodology. In the case of the A/V presentation, a virtualized MUSHRA GUI needs to be presented to the user as part of the visual scene.

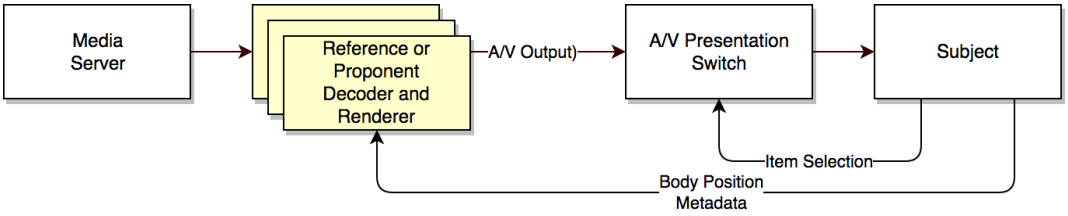

Figure 4 - Using MUSHRA-style near-instantaneous switching for VR/6DoF evaluation.

Output Documents
N17176 - MPEG-I Audio Use Cases
N17177 - MPEG-I Audio Architecture and Evaluation Procedures for 6 DoF
N17253 - Thoughts on Test Material and Encoder Input Format
N17254 - Workplan for MPEG-I 6 DoF Audio

5 Internet of Media Things

Architecture
The global IoMT interface is presented in the Figure below, which identifies a set of interfaces, protocols and associated media-related information representations related to:

- User commands (setup info.) between a system manager and an MThing, cf. Interface 1.
- User commands (Setup info.) forwarded by an MThing to another MThing, possibly in a modified form (e.g., subset of 1), cf. Interface 1'.
- Sensed data (Raw or processed data) (compressed or semantic extraction) and actuation information, cf. Interface 2.
- Wrapped interface 2 (e.g. for transmission), cf. Interface 2'
- MThing characteristics, discovery, cf. Interface 3.

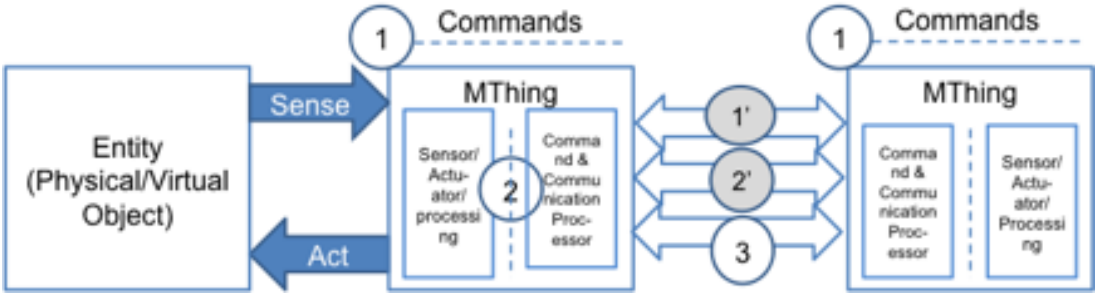

Figure 5 - Internet of Media Things Architecture

Use cases
The 23 overall IoMT use-cases are structured in four main categories, as follows:

- **Smart spaces: Monitoring and control with network of audio-video cameras**
  - Face recognition to evoke sensorial actuations
  - Human tracking with multiple network cameras
  - Networked digital signs for customized advertisement
  - Intelligent firefighting with IP surveillance cameras
  - Automatic video clip generation by detecting event information
  - Self-adaptive quality of experience for multimedia applications
  - Ultra wide viewing video composition
  - Digital signage and second screen use
  - Temporal synchronization of multiple videos for creating 360° or Multiple view video
- **Smart spaces: Navigation**
  - Blind person assistant system
  - Personalized navigation by visual communication
  - Personalized tourist navigation with natural language functionalities
  - Smart identifier: face Recognition on Smart Glasses
  - Smart advertisement: QR code recognition on smart glasses

- **Smart environments in smart cities**
    - Smart factory: Car maintenance assistance A/V system using smart glasses
    - Smart museum: Augmented visit museum using smart glasses
    - Smart house: Light control, vibrating subtitle, olfaction media content consumption
    - Smart car: Head-light adjustment and speed monitoring to provide automatic volume control
- **Smart collaborative health**
    - Increasing patient autonomy by remote control of left-ventricular assisted devices
    - Diabetic coma prevention by monitoring networks of in-body / near body sensors
    - Enhanced physical activity with smart fabrics networks
    - Medical assistance with smart glasses

Output Documents
N17226 - WD 3.0 of IoMT Part 1 Architecture
N17227 - WD 3.0 of IoMT Part 2 Discovery and Communication API
N17228 - WD 3.0 of IoMT Part 3 Media data formats and API