International Broadcasting Convention 2017 (IBC2017)
14-18 September 2017

Media orchestration between streams and devices via new MPEG timed metadata

IET Journals
The Institution of Engineering and Technology
ISSN 2515-236X
doi: 10.1049/oap-ibc.2017.0322
www.ietdl.org

M. Oskar van Deventer[1] ✉, Jean-Claude Dufourd[2], Sejin Oh[3], Seong Yong Lim[4], Youngkwon Lim[5], Krishna Chandramouli[6], Rob Koenen[1,7]

[1]TNO, Den Haag, Netherlands
[2]Telecom Paristech, Paris, France
[3]LGE, Seoul, Korea
[4]ETRI, Seoul, Korea
[5]Samsung, Seoul, Korea
[6]QMUL, London, UK
[7]tiledMedia, Rotterdam, Netherlands
✉ E-mail: oskar.vandeventer@tno.nl

Abstract: The proliferation of new capabilities in affordable smart devices capable of capturing, processing and rendering audio-visual media content triggers a need for coordination and orchestration between these devices and their capabilities, and of the content flowing from and to such devices. The upcoming Moving Picture Experts Group (MPEG) Media Orchestration ('MORE', ISO/IEC 23001-13) standard enables the temporal and spatial orchestration of multiple media and metadata streams. Temporal orchestration is about time synchronisation of media and sensor captures, processing and renderings, for which the MORE standard uses and extends a DVB standard. Spatial orchestration is about the alignment of (global) position, altitude and orientation, for which the MORE standard provides dedicated timed metadata. Other types of orchestration involve timed metadata for region of interest, perceptual quality of media, audio-feature extraction and media timeline correlation. This study presents the status of the MORE standard, as well as associated technical and experimental support materials. The authors also link MORE to the recently initiated MPEG-I (MPEG Immersive) project.

1 Introduction

A typical household may own more than ten internet-connected media devices, including smart TVs, tablet devices, smartphones and smartwatches. The combined use of devices may enhance the media consumption experience; for example, the recent HbbTV 2.0 standard [1] enables a smart TV to be connected to a tablet device for new types of interactive and synchronised media applications. New opportunities for media orchestration also arise at the capture side, as the number of cameras, microphones and sensors (location, orientation) may match the number of people present at large sports, music or other events. Moreover, the emergence of 360° video ('virtual reality') and associated 3D audio offer opportunities for less TV-centric orchestrations of media capture, processing and rendering.

Moving Picture Experts Group (MPEG) initiated its Media Orchestration ('MORE') activity early 2015 to create tools to manage multiple, heterogeneous devices over multiple, heterogeneous networks, orchestrating the devices, media streams and resources to create a single media experience. The focus of the activity has been on temporal and spatial orchestration, that is, protocols and metadata for the time synchronisation of media capture and media renderings, as well as their spatial alignment. The work has resulted in a draft international standard [2], which is expected to be formally published early 2018.

The remainder of this paper discusses use cases for media orchestration, details of the functional architecture and associated metadata, and a discussion how media orchestration fits in the MPEG Immersive (MPEG-I) roadmap.

2 Use cases for media orchestration

MPEG builds its standards on a set of requirements, which are typically derived from a set of use cases. MPEG collected a large number of use cases that require the orchestration of media devices, and then combined these to have a set of distinct use cases that were clearly different, requiring clearly different functionality. The use cases in the following are adapted from these.

2.1 Advanced multi-camera video stitching

The demand for wide-field-of-view video is increasing to provide immersiveness. It requires panoramic videos with wide horizontal and vertical angles. Fig. 1 presents a simple comparison between several video formats. Even though it depends on shooting environments, multiple camera systems are usually able to provide wide-viewing angles which are enough to cover the whole human vision area. For that purpose, there are several specific solutions to keep the pixel density and a distortion-free view.

Fig. 2 shows two types of multiple camera systems and a real-time monitoring system. The use of multiple cameras reduces radial distortion caused by wide-viewing-angle lenses. However, it requires a stitching process supported by high-performance GPUs to produce the seamless video with multiple video streams in real time. This complicated process requires media orchestration in the form of spatial information of multiple cameras and target objects, as well as temporally synchronised video streams.

2.2 Tracking persons of interest over multiple street views

In the security domain, investigators are required to construct an event narrative, by stitching together a single video stream obtained from multiple cameras of different types, see Fig. 3. The orchestration methodology for combining CCTV footage with different types of user content requires spatial and temporal, media orchestration based on distinctive regions or patterns (DROP) [3].

Fig. 1 *Immersive experience from wide view angle video*

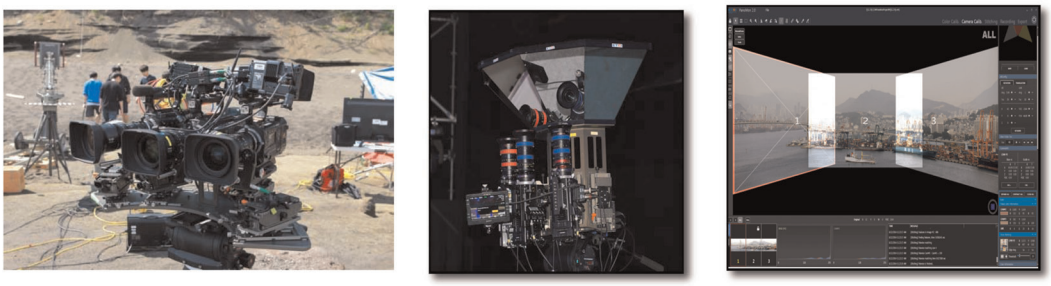

Fig. 2 *Multiple camera systems and real-time monitoring system for panoramic video*

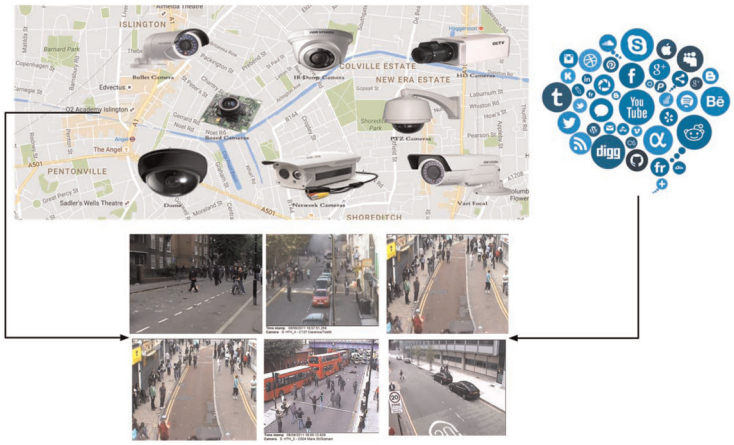

Fig. 3 *Tracking people over multiple street views*

In DROP-based media orchestration, each captured video stream has its own timeline, different and independent of the other video streams. The use cases require correlation of timelines between the different video streams. The media orchestration specification provides architectural components and data formats to enable the provision of universal media timeline in which heterogeneous timed media can be represented from disparate sources.

2.3 Virtual reality-related media orchestration use case

Multiple resources are used to generate virtual reality and augmented reality contents. Multiple cameras are required to capture 360° and in addition to the cameras, graphical contents are required to generate augmented reality contents. The Media Orchestration specification will provide technologies to efficiently implement such use cases.

One of the interesting use cases along this line is generating content covering 360° with a large number of uncoordinated video feeds. On the location of a concert/festival, there are usually a number of video feeds, both professional (good quality, reliable, continuous) and amateur (any quality, unreliable, on and off randomly), as well as picture contributions (of all qualities) to the capture scene. A (e.g. distant) user will have a means to browse the capture scene and select a point of view dynamically. The system implementing such use case requires to have means to dynamically orchestrate the media resources and stitch input video feeds according to the view selected by a user. In some cases, a video corresponding to a synthetic point of view only partially covered by some video feeds, in which case a combination of photos and videos dynamically selected are used to construct the synthetic point of view, will be generated.

Immersive Coverage of Spatially Outspread Live Events (ICoSOLE) is a project developing a system for such use case [4]. ICoSOLE aims at supporting use cases that enable users to experience live events which are spatially spread out, such as festivals (e.g. Gentse feesten in Belgium, Glastonbury in the UK), parades, marathons or bike races, in an immersive way by combining high-quality spatial video and audio and user-generated content.

3 Architecture for media orchestration

Media orchestration is about capture and consumption of many media with the help of (timed) metadata. The scope of the specification includes any combination of media and metadata to produce more media and metadata: synchronisation, stitching/

IBC, Open Access Proc. J., 2017, pp. 1–4

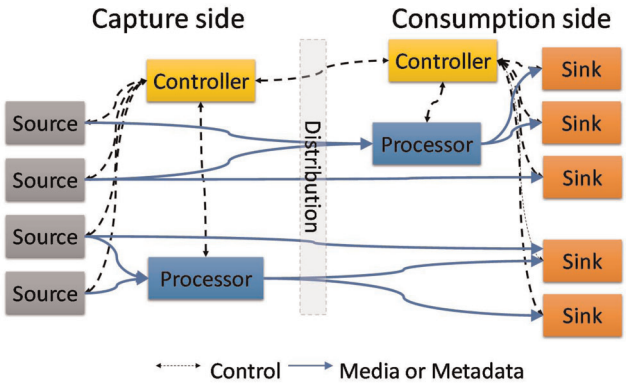

Fig. 4 *Architecture of media orchestration*

mixing, composition and more. Capture and consumption can be mixed quite intricately in some applications but conceptually they are separate.

Fig. 4 illustrates the media orchestration architecture. On the capture side, there are 'sources': sources of media and/or sources of metadata. On the consumption side, there are 'sinks', which present media to consumers according to metadata and orchestration information. On both sides, processors may transform media and/or metadata, and controllers create orchestration information and control the orchestration through messaging.

Relevant media types include audio, video, 3D, 360 video etc. Relevant metadata types include anything from synchronisation metadata (spatial, temporal) to semantic information, capture location and direction to object position tracking in a media.

3.1 Functional elements

This section describes the roles of functional elements of Media Orchestration.

A source captures media and/or creates metadata and is capable to stream them. A video camera is a source of media. A GPS is a source of location metadata. A source registers with a controller and then listens to its messages: such messages include to start or stop the capture, to configure capture parameters including encoding, format, quality, focal length, capture direction etc.

A sink presents one or more media to consumers, possibly driven by metadata such as composition metadata. Multiple media maybe presented on one sink, which is enabled by scene or composition metadata. Multiple sinks may be used to present a single media or a synchronised presentation with a help of a particular processor.

A processor transforms media and/or metadata into media and/or metadata. A transcoder from a video format to another is a possible media-to-media processor. An analyser generating football player positions from a video stream is an example of a media-to-metadata processor. A video stitcher, transforming multiple videos into one 360 video, is another example of processor.

A controller keeps a list of available sources, sinks and processors; and it organises connections, distribution, storage and retrieval of data. One controller can manage both capture and consumption for some applications where both happen at the same time; or there can be a controller (or more) on each side, for example if consumption happens later.

In Fig. 4, the capture controller registers the sources, connects them to the processor inputs and to distribution, triggers capture and processing and passes information to the consumption controller. The consumption controller registers the sinks and processors, connects processor outputs and distribution to sinks. Controllers manage new sources and sinks or their variable availability, locations, quality. The difference between a controller and a processor is that a controller receives and sends messages dealing with the set-up of media orchestration, whereas a processor deals with data used for the consumer experience, either directly media, or data used to be able to present media, such as synchronisation or composition information.

A media orchestration application (outside of the scope of the standard) would run on top of the controllers. Such an application could enable editing by a human director, presenting all the information available to the controllers in a concise manner, and relaying directions. The MPEG MORE standard is planned to include APIs allowing the application to make use of web technologies for such a user interface.

3.2 Temporal orchestration

The MPEG MORE standard reuses and extends the DVB Companion Streams and Screens (CSS) standard [5] for temporal orchestration. DVB CSS specifies a set of protocols that enable media synchronisation of a media stream on a TV and one on a tablet. DVB CSS is a TV-centric specification, focusing on e.g. alternative audio or ancillary video played on a tablet along the main video stream shown on the big TV screen, see also [6].

Two DVB CSS protocols are reused by MPEG MORE: the Wall Clock (WC) protocol and the Timeline Synchronisation (TS) protocol, see Fig. 5. The WC protocol creates a uniform and consistent reference clock on all the synchronised devices. The term 'wall clock' is a bit of a misnomer, as the wall clock time is unrelated to UTC or local time. The TS protocol coordinates the wall clock times at which identified video frames or audio samples should be presented to the user on the different devices. The devices each report the earliest wall-clock time that they could present an identified video frame or audio sample and coordinate playout delays accordingly. Video frames and audio samples are identified by timestamps in their media container, e.g. composition timestamp (ISO Based Media File Format) or presentation timestamp [MPEG-2 transport stream (TS)].

MPEG MORE extends various aspects of DVB CSS.

Whereas DVB CSS focuses on timed media data, MPEG MORE also includes timed metadata. An example is the TS between a video stream and a stand-alone position stream, e.g. a timed stream of GPS data. As timed media data and timed metadata are carried in the same types of containers, DVB CSS can be reused as is for timed metadata.

Whereas DVB CSS covers only render/sink-side TS, MPEG MORE also covers capture/source-side TS. The TS protocol is extended for this purpose, coordinating the wall clock time at which identified video frames or audio samples are captured.

A third aspect is timeline correlation. If different timed media or timed metadata have different time bases, then it is needed to know the clock skew, clock drift and clock-drift variation between the two time bases. As the DVB CSS solution for this was too TV centric, MPEG MORE specifies its own solution, namely metadata for timeline correlation.

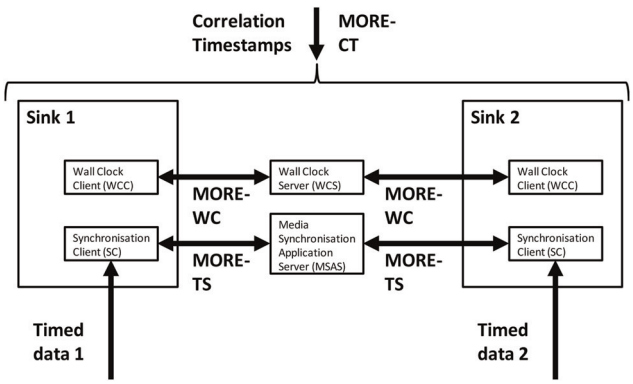

Fig. 5 *Architecture for temporal orchestration [2]*

3.3 Spatial orchestration

MPEG MORE considers use cases with multiple sources as well as sinks including one-to-many and many-to-one scenarios. One or more streams from multiple sources are dynamically played out across multiple sinks. These multiple media streams can be combined to provide a single immersive experience, e.g. omnidirectional video.

In order to achieve this, MPEG MORE considers how to arrange and coordinate media streams based on its spatial relationships when the location and orientation (gaze) of sources are tracked in a 3D environment. For example, when a tracker (that is capable of tracking location and orientation) is attached to a camera, the location and orientation of the camera are continuously tracked as the video is being captured. With a captured video stream, the position and orientation streams of the camera are also generated. The position and orientation streams are timed metadata about the associated video stream according to its intrinsic timeline. Both media and metadata streams are delivered to the sinks, or they can be used by processors that exist between sources and sinks. For spatial orchestration of multiple media streams generated from independent sources, the position and orientation streams are used to recognise the spatial relationship of associated media streams and to coordinate the media streams. Metadata can also be used for supporting dynamic media presentation according to the sink's movement.

3.4 Carriage of timed metadata for media orchestration

Media orchestration metadata is defined as data that cannot be rendered independently and may affect rendering, processing or orchestration of the associated media data. Like media data, this metadata can be timed metadata which has an intrinsic timeline.

For supporting temporal or spatial orchestration of multiple media streams, MPEG MORE describes several types of timed metadata, e.g. position, altitude, orientation, quality, stream monitor, and specifies how to carry the timed metadata in ISO Base Media File Format (ISOBMFF) or MPEG-2 TSs.

The carriage of timed metadata in ISOBMFF is useful in cases where media data and associated timed metadata are stored in files, either together or in separate files. In this case, the timed metadata is carried in the metadata tracks within an ISOBMFF file. Different metadata types and corresponding storage formats are identified by their unique sample entry codes.

The carriage of timed metadata in MPEG-2 TS (transport stream) is useful in cases of broadcast of media data and associated timed metadata. The timed metadata associated with one or more video or audio frame are stored in access units and encapsulated in a PES stream.

4 Media orchestration and immersive media

MPEG's MORE activity is related to the recently initiated MPEG-I project. MPEG-I sets a set of standards for the Coded Representation of Immersive Media. The goals of MPEG-I are the following, paraphrased from MPEG-internal documents.

New devices and services emerge that allow users to be immersed in media, and navigate multimedia scenes. A fragmented market exists for such devices and services, notably for content that is delivered 'over the top'. This is because no common standards exist for the representation and delivery of such content and services.

MPEG-I seeks to provide such standards, to enable interoperable services and devices that provide immersive, navigable experiences. MPEG-I seeks to enable the types of services that are available today, as well as to support the evolution in immersive media that is expected to continue for the foreseeable future.

There is a close relation between what MPEG MORE provides and the technologies needed for the type of immersive services that MPEG-I targets. For example, some of MPEG MORE's use cases concern many-camera systems, where the service seeks to immerse users in the output of those sources. MPEG MORE provides tools that allow the orchestration of sources, to combine those sources into a single immersive experience. The resulting experience could play in a head-mounted device, but it could also be reproduced on a number of distinct 'sinks' – usually screens and speakers, and devices that combine these. MORE allows the spatial and temporal coordination between these devices, for a harmonised and immersive experience. We, therefore, expect that MPEG MORE will be a useful and important specification for MPEG-I to use and reference.

5 Conclusions

This paper introduced the MPEG draft specification on MORE. This specification provides metadata and protocols for temporal and spatial orchestration between multiple media capture and rendering devices. Use cases include advanced multi-camera video stitching, tracking persons of interest over multiple street views as well as immersive media. The specification provides an architecture with sources, sinks, processors and controllers, as well as audio-visual media data, metadata and orchestration data. Temporal orchestration (synchronisation of media data and metadata) is achieved by reusing and extending protocols from DVB CSS. Spatial orchestration is achieved with new timed location and orientation metadata that is associated with the audio-visual media data. The new timed metadata is specified to be carried in ISOBMFF files, MPEG-2 TSs, as well as MPEG DASH and MPEG MMT transport. It is expected that media orchestration will play a role in the production and consumption of immersive media, as part of the MPEG-I project.

6 References

1 HbbTV: 'Hybrid broadcast broadband television version 2.0', ETSI TS 102 796 V1.4.1, 2016. Available at http://www.etsi.org/deliver/etsi_ts/102700_102799/102796

2 MPEG: 'Media Orchestration' (MORE), ISO/IEC 23001-13:2018, draft available via authors

3 LASIE: 'Large-scale information exploitation of forensic data'. Available at http://www.lasie-project.eu

4 ICoSOLE: 'Immersive coverage of spatially outspread live events'. Available at http://icosole.eu/public-deliverables

5 DVB: 'Companion streams and streams', ETSI TS 103 286 02 v1.1.1. Available at http://www.etsi.org/deliver/etsi_ts/103200_103299/10328602, 2015

6 Oskar van Deventer, M., Stokking, H., Hammond, M., *et al.*: 'Standards for multi-stream and multi-device media synchronization', IEEE Commun. Mag. Commun. Stand. Suppl., 2016, **54**, (3), pp. 16–21