

Media Orchestration Between Streams and Devices via New MPEG Timed Metadata

By M. Oskar van Deventer, Jean-Claude Dufourd, Sejin Oh, Seong Yong Lim, Youngkwon Lim, Krishna Chandramouli, and Rob Koenen

Abstract

The proliferation of new capabilities in affordable smart devices capable of capturing, processing, and rendering audiovisual 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 standard (MORE, ISO/IEC 23001-13) enables the temporal and spatial orchestration of multiple media and metadata streams. The temporal orchestration is about time synchronization of media and sensor captures, processing, and renderings, for which the MORE standard uses and extends a Digital Video Broadcasting standard. The 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 the region of interest, perceptual quality of media, audio-feature extraction, and media timeline correlation. This paper presents the status of the MORE standard as well as the associated technical and experimental support materials. We also link MORE to the recently initiated MPEG immersive project.

This paper presents the status of the MORE standard as well as the associated technical and experimental support materials.

Keywords

Broadcast, ISO, media, media orchestration, MPEG, MPEG-I, MPEG MORE, multimedia, spatial, television, temporal

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 standard¹ enables a smart TV to be connected to a tablet device for new types of interactive and synchronized media applications. New opportunities for media orchestration (MORE) also arise at the

capture side, as the number of cameras, microphones, and sensors (location and orientation) may match the number of people present at large sports, music, or other events. Moreover, the emergence of 360° video [virtual reality (VR)] and the associated 3D audio offer opportunities for less TV-centric orchestrations of media capture, processing, and rendering.

The Moving Picture Experts Group (MPEG) initiated its MORE activity in 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, i.e., protocols and metadata for the time synchronization of media capture and media renderings, as well as their spatial alignment. The work has resulted in a draft international standard that is expected to be formally published at the end of 2018 or early 2019. The remainder of this paper discusses use cases for MORE, details of the functional architecture and the associated metadata, and how MORE fits in the MPEG-immersive (MPEG-I) roadmap.

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 functionalities. The use cases below are adapted from these.

Advanced Multicamera 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. **Figure 1** presents a simple comparison between several video formats. Even though it depends on shooting environments, multiple camera systems are usually able to

Digital Object Identifier 10.5594/JMI.2018.2870019
Date of publication: 9 November 2018



FIGURE 1. Immersive experience from wide-view-angle video.

provide wide viewing angles that 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.

Figure 2 shows two types of multiple camera systems and a realtime 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 Graphics Processing Units to produce a seamless video with multiple video streams in realtime. This complicated process requires MORE in the form of spatial information of multiple cameras and target objects as well as temporally synchronized video streams.

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 (Fig. 3). The orchestration methodology for combining CCTV footage with different types of user content requires spatial and temporal MORE based on distinctive regions or patterns (DROP).² In DROP-based MORE, each captured video stream has its own timeline, different and independent of the other video streams. The use cases require a correlation of timelines between the different video streams. The MORE specification provides architectural components and data formats to enable the provision of a universal media timeline in which heterogeneous timed media can be represented from disparate sources.

Virtual Reality Related Media Orchestration Use Case

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

An interesting use case 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, and continuous) and amateur (any quality, unreliable, on and off randomly), as well as picture contributions (of all qualities) to the captured scene. A (e.g., distant) user will have the means to browse the capture scene and select a point of view dynamically. The system implementing such a use case requires 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 photographs and dynamically selected videos 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 a use case.³ ICoSOLE aims at supporting use cases that enable users to experience live events that are spatially spread out, such as festivals (e.g., Gentse feesten in Belgium, Glastonbury, in the U.K.), parades,

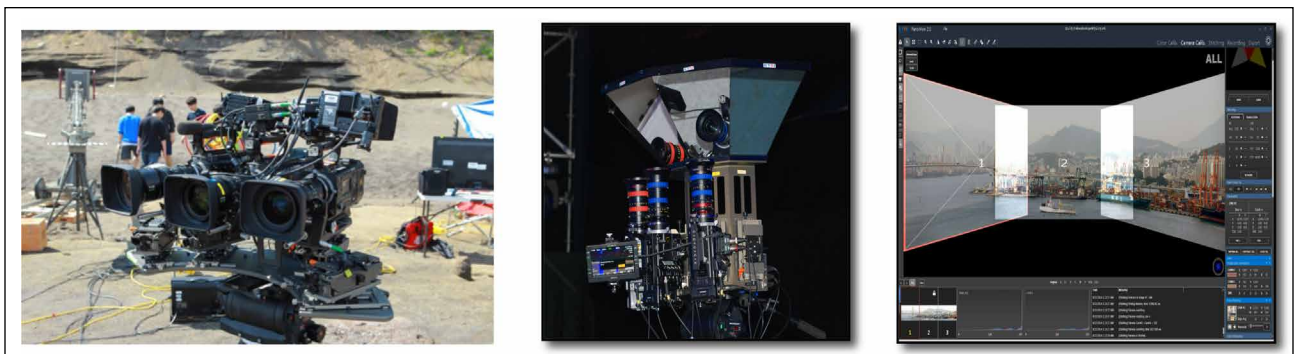


FIGURE 2. Multiple camera systems and realtime monitoring system for panoramic video.

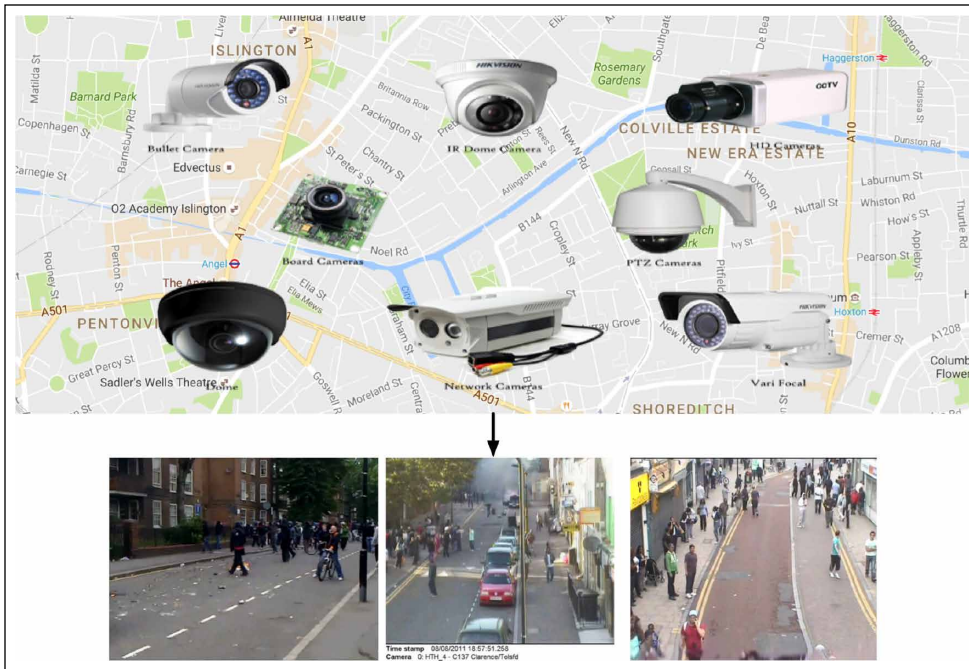


FIGURE 3. Tracking people over multiple street views.

marathons, or bike races, in an immersive way by combining high-quality spatial video and audio and user-generated content.

Architecture for Media Orchestration

MORE 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, i.e., synchronization, stitching/mixing, composition, and more. Capture and consumption can be mixed quite intricately in some applications, but conceptually they are separate.

Figure 4 illustrates the MORE architecture. On the capture side, there are sources, i.e., sources of media and/or sources of metadata. On the consumption side, there are sinks that 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 and 360 video. Relevant metadata types include anything from synchronization metadata (spatial and temporal) to semantic information, capture location, and direction to object position tracking in a media.

Functional Elements

This section describes the roles of functional elements of MORE. A source captures media and/or creates metadata and is capable of streaming 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 as to start or stop the capture

or 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 may be presented on one sink, which is enabled by scene or composition metadata. Multiple sinks may be used to present a single media or a synchronized presentation with the 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 analyzer 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 a processor.

A controller keeps a list of available sources, sinks, and processors; it also organizes connections, distribution, storage, and retrieval of data. One controller can

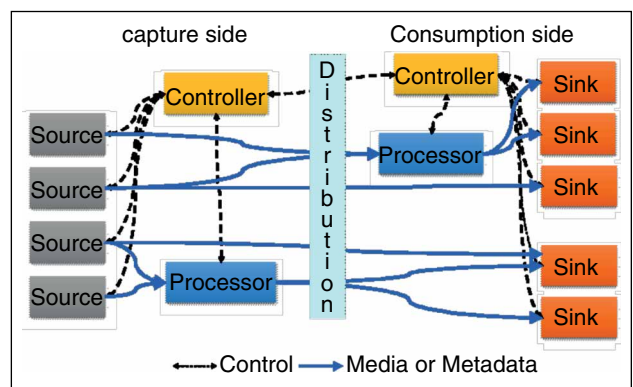


FIGURE 4. Architecture of MORE.

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, and connects processor outputs and distribution to sinks. Controllers manage new sources and sinks or their variable availability, locations, and quality. The difference between a controller and a processor is that a controller receives and sends messages dealing with the setup of MORE, whereas a processor deals with data used for the consumer experience, either directly media or data used to present media, such as synchronization or composition information.

A MORE 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 Application Programming Interfaces allowing the application to make use of web technologies for such a user interface.

Temporal Orchestration

The MPEG-MORE standard reuses and extends the Digital Video Broadcasting (DVB) companion streams and screens (CSS) standard⁴ for temporal orchestration. DVB-CSS specifies a set of protocols that enable media synchronization 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.⁵

Two DVB-CSS protocols are reused by MPEG-MORE: the wall clock (WC) protocol and the timeline synchronization (TS) protocol (**Fig. 5**). The WC protocol creates a uniform and consistent reference clock on all the synchronized devices. The term WC is a bit of a misnomer, as the WC time is unrelated to Coordinated Universal Time or local time. The TS protocol coordinates the WC 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 WC 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 (ISOBMFF)] or presentation timestamp (MPEG-2 Transport Stream).

MPEG-MORE extends various aspects of DVB-CSS. While DVB-CSS focuses on timed media data, MPEG-MORE also includes timed metadata. An example is the timeline synchronization between a video stream and a standalone 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.

While DVB-CSS covers only render/sink-side timeline synchronization, MPEG-MORE also covers capture/source-side timeline synchronization. The TS protocol is extended for this purpose, coordinating the WC 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 necessary 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.

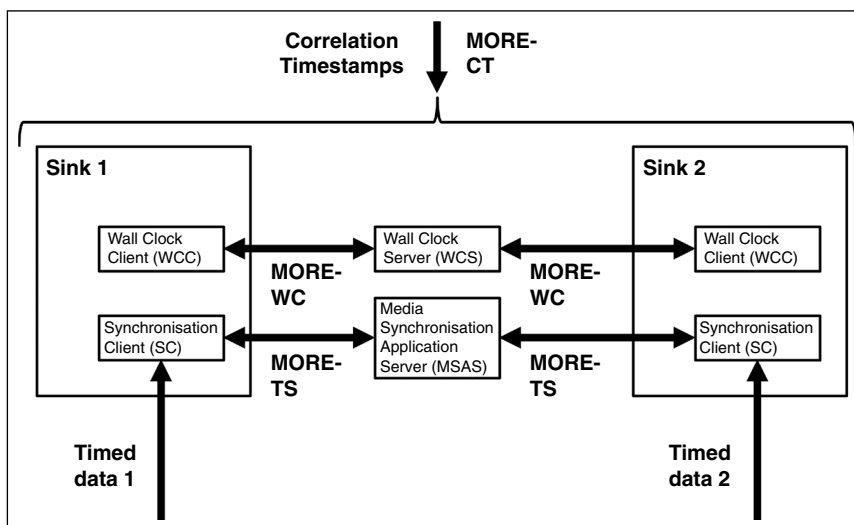


FIGURE 5. Architecture for temporal orchestration.⁶

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 able 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 recognize 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.

Carriage of Timed Metadata for Media Orchestration

MORE 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 that have an intrinsic timeline.

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

The carriage of timed metadata in ISOBMFF is useful in cases where media data and the 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 the corresponding storage formats are identified by their unique sample entry codes.

The carriage of timed metadata in the MPEG-2 transport stream is useful in cases of the 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 Packetized Elementary Stream stream.

Media Orchestration and Immersive Media

MPEG's MORE activity is related to the MPEG-I project, which was initiated in 2016. MPEG-I sets

a set of standards for the Coded Representation of Immersive Media. The goals of MPEG-I are as follows, paraphrased from MPEG-internal documents.⁷

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 of immersive media that is expected to continue for the foreseeable future.

There is a close relationship 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. MPEG-MORE allows the spatial and temporal coordination between these devices for a harmonized and immersive experience. We therefore expect that MPEG-MORE will be a useful and important specification for MPEG-I to use and reference.

Conclusion

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 multicamera video stitching and 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 audiovisual media data, metadata, and orchestration data. Temporal orchestration (synchronization of media data and metadata) is achieved by reusing and extending protocols from DVB-CSS. The spatial orchestration is achieved with new timed location and orientation metadata that is associated with the audiovisual media data. The new timed metadata are specified to be carried in ISOBMFF files, MPEG-2 Transport streams, as well as MPEG-DASH and MPEG media transport. It is expected that MORE will play a role in the production and consumption of immersive media as part of the MPEG-I project.

References

1. Hybrid Broadcast Broadband Television (HbbTV), “HbbTV version 2.0,” ETSI TS 102 796 V1.4.1, 2016. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/102700_102799/102796/01.04.01_60/ts_102796v010401p.pdf
2. Large-Scale Information Exploitation of Forensic Data (LASIE), [Online]. Available: <http://www.lasie-project.eu>
3. Immersive Coverage of Spatially Outspread Live Events (ICoSOLE), [Online]. Available: <http://icosole.eu/public-deliverables>
4. Digital Video Broadcasting (DVB), “Companion Streams and Streams,” ETSI TS 103 286 02 v1.1.1, 2015. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/103200_103299/10328602/01.01_60/ts_10328602v010101p.pdf
5. M. O. van Deventer, H. Stokking, M. Hammond, J. Le Feuvre, and P. Cesar, “Standards for Multi-Stream and Multi-Device Media Synchronization,” *IEEE Comm. Mag. Comm. Stand. Supp.*, 16–21, Mar. 2016.
6. Motion Picture Experts Group (MPEG), “Media Orchestration (MORE),” ISO/IEC 23001-13:2018, draft available via authors.
7. Motion Picture Experts Group (MPEG), “MPEG-I, Coded Representation of Immersive Media.” [Online]. Available: <https://mpeg.chiariglione.org/standards/mpeg-i>

About the Authors



M. Oskar van Deventer is a senior scientist in media networking. His focus is on international standards. He is one of the main contributors to the Motion Picture Experts Group (MPEG) MORE standard on media orchestration. He has been an editor of *DVB Companion Screens and Streams* and an active contributor to Hybrid broadcast broadband TV (HbbTV), European Telecommunication Standards Institute, Internet Engineering Task Force, Open Internet Protocol TeleVision Forum, and MPEG in the areas of media synchronization, content delivery network interconnection, adaptive streaming, and MORE. He was leading a work-package in the European HBB-Next R&D project on the next-generation hybrid broadcast-broadband television, where he developed technologies for media synchronization and group recommendations of media content. He has won several international awards in the area of mobile gaming. He is a co-author of one book, more than 150 publications, more than 80 patent applications, over 750 standardization contributions, and holds a Guinness world record.



Jean-Claude Dufourd has been a professor at Telecom ParisTech (TPT) since 2002. Both a teacher and researcher at TPT since 1990, he collaborated on several European and industry projects related to the Motion Picture Experts Group (MPEG)-4 and MPEG-21 standards. He also co-founded and was chief scientist of the company Streamazzo from 2004 to 2008 and co-founded the company MotionTree in 2012. His activities span the multimedia domain, from standards

to implementation, playing to authoring, and client to server. His team develops and maintains the GPAC Project on Advanced Content multimedia framework. A contributor to the MPEG standard since 1995, he was chair of the MPEG Integration Group and was responsible for coordinating compliance and reference software issues from 2002 to 2006. He also contributed to W3C, Open Mobile Alliance, 3GPP, DVB, and HbbTV. He is a former student of the Ecole Normale Supérieure in Paris, a graduate from TPT, and holds a PhD in computer science.



Sejin Oh is a chief research engineer in The Advanced Standard Research and Development Laboratory at LG Electronics Inc. Since she joined LG Electronics in 2011, she has been involved in the development and standardization of support immersive media systems. Her areas of expertise are media standardization (Motion Picture Experts Group, 3GPP, Advanced Television Systems Committee (ATSC), Digital Video Broadcasting (DVB), and Consumer Technology Association (CTA)), especially the video and audio aspect of virtual reality/augmented reality systems. She has generated more than 50 patents with relevance to standards. Oh received a PhD in immersive and intelligent media in augmented reality systems from the Gwangju Institute of Science and Technology, South Korea. Her research is focused on developing various kinds of immersive systems from immersive display systems to smartphones.



Seong Yong Lim received BS and MS degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 1999 and 2011, respectively. He is currently a PhD candidate with the Department of Electrical Engineering, KAIST. He joined the Electronics and Telecommunications Research Institute in 2001, where he is a principal researcher. His research interests include discrete event system, defense system modeling, and simulation.



Youngkwon Lim received BS and MS degrees in electronics engineering from Korean Aerospace University, Seoul, South Korea, in 1994 and 1996, respectively, and a PhD degree in electronics engineering from Hanyang University, Seoul, South Korea, in 2011. After he joined the Electronics and Telecommunications Research Institute, Daejeon, South Korea, in 1996, he actively participated in the

standardization and development of MPEG technologies by serving as chair on a number of ad hoc groups and editor of various MPEG standards. In 2000, he joined net&tv Inc. and has been leading the development of interactive broadcasting service solutions using Digital Multimedia Broadcasting standards. Since 2013, he joined Samsung Research America and has been working on the multimedia delivery system and its application in broadcasting services and mobile network services. He is currently serving as chair of the Systems Subgroup of MPEG, a role that he has been serving since 2009. He is also serving as chair of the ATSC Specialist Group Management and Protocols, which is developing ATSC 3.0 standards. His research interests include multimedia systems and convergence between digital broadcasting and the internet. He is a Member of the IEEE.




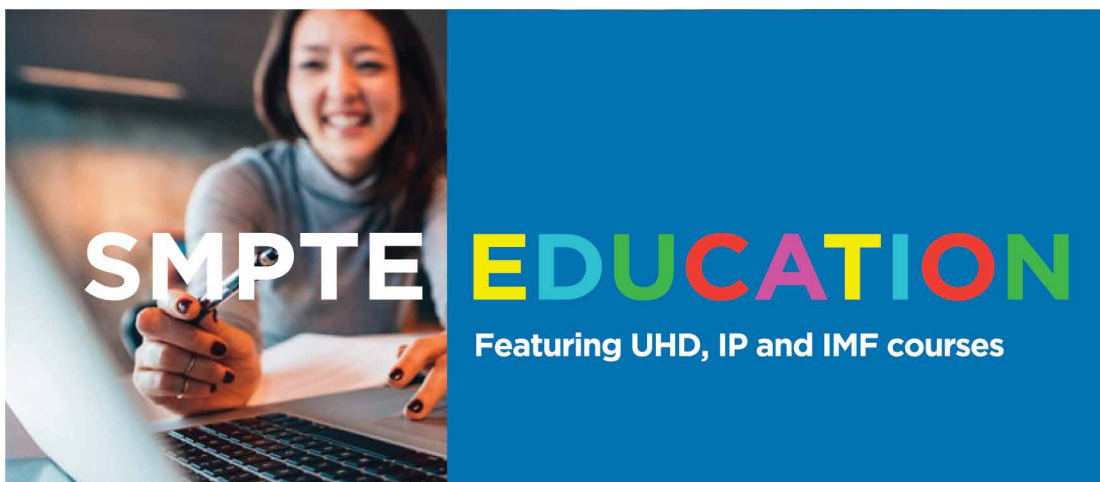
Krishna Chandramouli jointly holds the positions of research assistant at Queen Mary, University of London, and chief technology officer at Venaka Media Ltd. His field of expertise includes big-data analytics, machine learning, video analytics, time-series analytics, and knowledge modeling. He serves as a recognized subject expert at British

Standards International for systems and contributes toward international standards development. He received a PhD from the University of London in 2009 and has published more than 50 peer-reviewed research articles at international conferences and journals. He is a Member of the IEEE, the IEEE Computer Society, and ACM.



Rob Koenen is co-founder and chief business officer of Tiledmedia, the leading tiled virtual reality (VR) streaming company. He is also a principal with the Media Networking Group of TNO, The Netherlands' largest independent research institute. He is president of the VR Industry Forum, a cross-industry initiative that seeks to enable interoperable, high-quality VR experiences for consumers. His activities include the development of multimedia technologies, systems, and standards, and their application in novel services. He received an MSEE (ingenieur) degree from the Delft University of Technology, The Netherlands, in 1989, where he studied electrical engineering, specializing in information theory.

Presented at IBC2017, Amsterdam, The Netherlands, 14–18 September 2017. This paper is published here by kind permission of the IBC. Copyright © IBC. 



VIRTUAL CLASSROOM

One of SMPTe's most innovative educational offerings is the Virtual Classroom program. SMPTe provides convenient, high-value learning opportunities to members and other individuals from around the world. SMPTe Virtual Classroom courses are "blended learning" courses that include both independent study and live, instructor-led coaching sessions that cover more complex topics and activities.



THE NEXT CENTURY

View the latest offerings online and register today!
www.smpte.org/courses