International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

H.430.2
(08/2018)

SERIES H: AUDIOVISUAL AND MULTIMEDIA SYSTEMS

Infrastructure of audiovisual services – Telepresence, immersive environments, virtual and extended reality

Architectural framework for immersive live experience (ILE) services

Recommendation ITU-T H.430.2

ITU-T

ITU-T H-SERIES RECOMMENDATIONS

AUDIOVISUAL AND MULTIMEDIA SYSTEMS

CHARACTERISTICS OF VISUAL TELEPHONE SYSTEMS	H.100–H.199
INFRASTRUCTURE OF AUDIOVISUAL SERVICES	
General	H.200–H.219
Transmission multiplexing and synchronization	H.220–H.229
Systems aspects	H.230–H.239
Communication procedures	H.240–H.259
Coding of moving video	H.260–H.279
Related systems aspects	H.280–H.299
Systems and terminal equipment for audiovisual services	H.300–H.349
Directory services architecture for audiovisual and multimedia services	H.350–H.359
Quality of service architecture for audiovisual and multimedia services	H.360–H.369
Telepresence, immersive environments, virtual and extended reality	**H.420–H.439**
Supplementary services for multimedia	H.450–H.499
MOBILITY AND COLLABORATION PROCEDURES	
Overview of Mobility and Collaboration, definitions, protocols and procedures	H.500–H.509
Mobility for H-Series multimedia systems and services	H.510–H.519
Mobile multimedia collaboration applications and services	H.520–H.529
Security for mobile multimedia systems and services	H.530–H.539
Security for mobile multimedia collaboration applications and services	H.540–H.549
VEHICULAR GATEWAYS AND INTELLIGENT TRANSPORTATION SYSTEMS (ITS)	
Architecture for vehicular gateways	H.550–H.559
Vehicular gateway interfaces	H.560–H.569
BROADBAND, TRIPLE-PLAY AND ADVANCED MULTIMEDIA SERVICES	
Broadband multimedia services over VDSL	H.610–H.619
Advanced multimedia services and applications	H.620–H.629
Ubiquitous sensor network applications and Internet of Things	H.640–H.649
IPTV MULTIMEDIA SERVICES AND APPLICATIONS FOR IPTV	
General aspects	H.700–H.719
IPTV terminal devices	H.720–H.729
IPTV middleware	H.730–H.739
IPTV application event handling	H.740–H.749
IPTV metadata	H.750–H.759
IPTV multimedia application frameworks	H.760–H.769
IPTV service discovery up to consumption	H.770–H.779
Digital Signage	H.780–H.789
E-HEALTH MULTIMEDIA SYSTEMS, SERVICES AND APPLICATIONS	
Personal health systems	H.810–H.819
Interoperability compliance testing of personal health systems (HRN, PAN, LAN, TAN and WAN)	H.820–H.859
Multimedia e-health data exchange services	H.860–H.869
Safe listening	H.870–H.879

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T H.430.2

Architectural framework for immersive live experience (ILE) services

Summary

Recommendation ITU-T H.430.2 identifies the general framework and high-level functional architecture for immersive live experience (ILE) services, in order to clarify ILE services and the general functions of ILE systems. This Recommendation also provides the general role model of ILE.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T H.430.2	2018-08-29	16	11.1002/1000/13667

Keywords

General functions, general role model, high-level architecture, immersive live experience.

---

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T H.430.2

Architectural framework for immersive live experience (ILE) services

1 Scope

This Recommendation identifies the architectural frameworks of immersive live experience (ILE) services, including the high-level functional architecture and provides some candidate technologies for the functions required for ILE. In addition, the general role model for providing ILE services is defined as event promoters at the source site, media transport providers in the transmission part and public and live viewing promoters at the presentation sites.

The scope of this Recommendation includes:

– High-level architecture for ILE

– Candidate technologies for functions required for ILE

– General role model for providing ILE services.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

None.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following term defined elsewhere:

3.1.1 immersive live experience (ILE) [b-ITU-T H.430.1]: A shared viewing experience which stimulates emotions within audiences at both the event site and at remote sites, as if the viewers at remote sites had wandered into a substantial event site and had actually watched the events taking place in front of them. This impression is due to high-realistic sensations provided by a combination of multimedia technologies such as sensorial information acquisition, media processing, media transport, media synchronization and media presentation.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

CODEC	Coder-Decoder
DANTE	Digital Audio Network Through Ethernet
DMX	Digital Multiplex
ILE	Immersive Live Experience

MMT	MPEG Media Transport
MPEG	Moving Picture Expert Group
WFS	Wave Field Synthesis

5 Conventions

In this Recommendation:

– The keywords "is required to" indicate a requirement which must be strictly followed and from which no deviation is permitted, if conformance to this Recommendation is to be claimed.

– The keywords "is recommended" indicate a requirement which is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

– The keywords "can optionally" indicate an optional requirement that is permissible, without implying any sense of being recommended. This term is not intended to imply that the vendor's implementation must provide the option, and the feature can be optionally enabled by the network operator/service provider. Rather, it means the vendor may optionally provide the feature and still claim conformance with this Recommendation.

– The keyword "functions" are defined as a collection of functionalities. It is represented by the following symbol in this Recommendation:

Functions

– The keyword "functional block" is defined as a group of functionalities that has not been further subdivided at the level of detail described in this Recommendation. It is represented by the following symbol in this Recommendation:

Functional Block

NOTE – In the future, other groups or other Recommendations may possibly further subdivide these functional blocks.

Frame borders of "functions" and "functional blocks" and relational lines among "functions" and "functional blocks" are drawn with solid lines or dashed lines. The solid lines mean required functionalities or relations, while the dashed lines mean optional functionalities or relations.

6 High-level architecture of ILE

6.1 High-level architecture on ILE systems

In order to provide ILE services with live content streaming, the architecture requires a capturing environment, a presentation, an ILE application and a synchronous media transport and transport layer. A high-level architecture of an ILE system is shown in Figure 1. The content aggregated by a capturing environment at a source site, which includes cameras, microphones and sensors with lighting information, is transmitted simultaneously for presentation at viewing sites. The content is processed by asset processing and a coder-decoder (CODEC) on a media processing block and its signalling is reformatted by spatial information processing and synchronous information processing on a signalling processing block. These blocks are located on an ILE application. At the viewing sites, the delivered content is reconstructed at presentation. It is displayed on projectors and displays in real size or similar size with reproduced audio sound including sound direction and reconstructed lighting thanks to received lighting information. By presenting these synchronously at viewing sites, users experience high-realistic sensations.

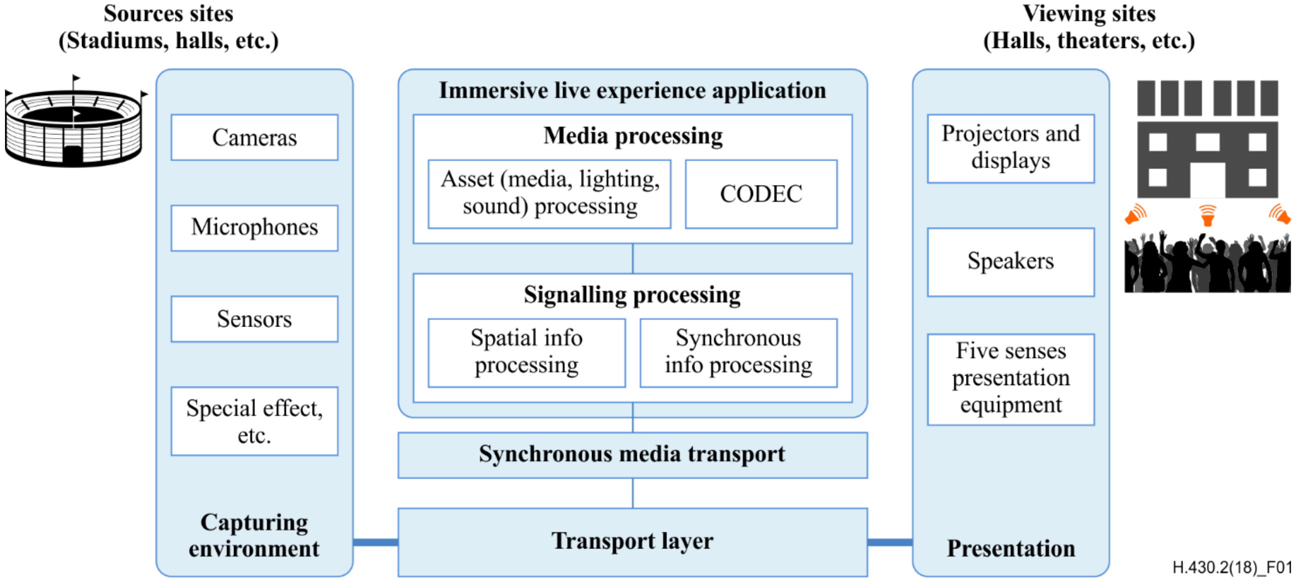

Figure 1 – A high-level architecture of immersive live experience system

6.2 Capturing environment on the source site

6.2.1 Cameras

In order to display pseudo 3D images at the viewing sites, it is required to capture stereoscopic images by using several cameras and to add spatial positions to captured images in consideration of real-sized displays. The cameras should be high-resolution as much as possible.

6.2.2 Microphones

Microphones are necessary to provide sound direction which means that the sounds should seem to be produced from objects at the viewing site. To achieve this, it is required to capture sounds from objects at the source site using multiple microphones and to add spatial positions to captured sounds for each object. The sound direction should be aligned with objects which are presented by display on terminal devices at the viewing site.

6.2.3 Sensors

To provide an immersive experience to participants at the viewing sites, it is best to provide environmental information, such as temperature and humidity, in addition to images and sounds. This kind of environmental information should be measured with various types of sensors. For example, vibrations at the source site may be transmitted and reproduced at viewing sites based on sensed environmental information.

Sensors are also used for presenting objects more realistically at the viewing sites. It is required to measure the 3D position of objects, such as players or musicians at the source site and to track these objects using combinations of several types of sensors, such as depth sensors.

6.2.4 Special effects and other stage effects

Usually events such as music concerts and other events comprise special effects or stage effects including lighting. Public viewing or live viewing sites are usually different from the source environment such as music concert halls and sports stadiums. In order to reconstruct the source site environment at the viewing sites, special effects or stage effects information including lighting information is beneficial for improving high-realistic sensations for users at the viewing sites.

6.3 Presentation functions on the viewing sites

6.3.1 Projectors and displays

There are several typical theatre styles, such as proscenium style, open style or amphitheatre/arena style. These various theatre styles are illustrated in Figure 2.

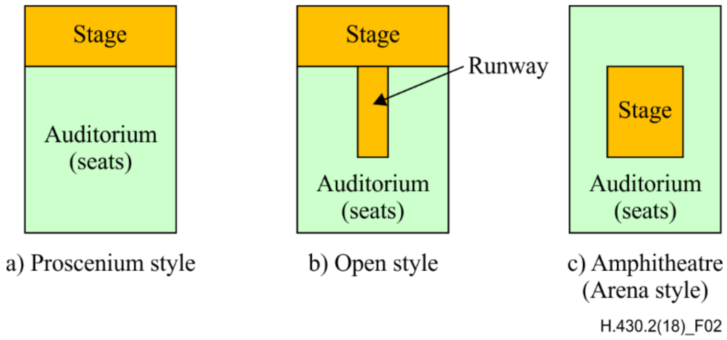

Figure 2 – Typical theatre styles

a) Proscenium style

In most theatres, audiences watch a stage or screens in front of them. As shown in Figure 2(a), a stage is located on one end of theatre and the seating area for the audience, called the auditorium, is located at the other end. This theatre style is called proscenium style.

Most viewing sites have a completely different environment in terms of for instance, hall size and capacity, from the source site. In addition, various types of display devices have a wide variety of specifications, i.e. a corresponding wide variety of resolutions and screen sizes. The spatial location of devices which includes horizontal and vertical positions and vent angles is different at each viewing site.

In order to reconstruct source objects and the environmental situation at the different viewing sites, spatial environment information including terminal device capabilities and their spatial location information are required. Furthermore, displaying real-sized objects also requires the spatial position, for example height from floor and width from the neighbouring display devices.

b) Open style

Many concert halls have one or more stages and/or runways, as shown in Figure 2(b), so audiences may watch from different directions. This type of theatre is called open style. Open style theatres vary widely. Some open style theatres have a stage with one or more runways while others have a stage and one or more island type stages.

For reconstructing the environment of a source site at viewing sites, it is required to consider layered projection or multi-screen projection in a 3D manner. This is considered as a way to provide spatial information of captured objects and environmental images from source sites to viewing sites.

c) Amphitheatre/Arena style

Amphitheatre/arena style theatres have an area or stage in the centre of the theatre and audiences watch from all around the stage, which means they can watch objects on the stage from anywhere in the 360-degree circumference, see Figure 2(c).

For audiences in amphitheatre/arena style viewing sites, pseudo-3D display systems are required to handle object positions to provide 3D-object viewing with 360-degree angles. 3D spatial projection enables display of 3D objects with smooth movement in the depth direction, without any head mounted displays.

6.3.2 Speakers

In order to reconstruct sound direction at viewing sites, the sound direction should be aligned with objects which are presented on displays by using sound data and spatial information captured from objects at the source site.

6.3.3 Five sense presentation equipment

Audiences at viewing sites want to feel the same atmosphere as the source site as much as possible, so sensed data from the source site should be reconstructed at the viewing sites.

In addition, special effects or stage effects information including lighting information sent from the source site could reproduce the atmosphere of the source site at the viewing sites to provide the ambience of the event venue. To make viewing sites more realistic, lighting at the viewing sites should have similar lighting equipment to that the source site and control it by lighting information from source site.

6.4 Functions in the media processing part

6.4.1 Media processing

The asset processing function in media processing block creates assets which include media frames by extracting and reconstructing content such as video and audio. CODEC in the media processing block encodes and decodes asset information.

6.4.2 Signalling processing

The spatial information function in the signalling block creates spatial information from the physical size and position of displays at viewing sites. Synchronous information processing reforms times of all content information to synchronize content such as video, audio, lighting and sensed information.

6.5 Functions in the transmission part

6.5.1 Synchronous media transport

The synchronous media transport function delivers several media synchronously. In order to achieve ILE services, especially when representing live pseudo 3D content and the reproduction of sound direction, media transport technologies such as MPEG media transport (MMT) with some enhancements and representation methods on terminal equipment should be considered. MMT technology standardized by ISO/IEC JTC1 SC29 WG11 (MPEG) [b-ISO/IEC 23008-1] has a capability for transmitting several media of asset information, but environmental information such as spatial information might not be transmitted in its original specification. Therefore, there may need to be an enhanced functional capability to MMT for presenting media with environmental information synchronously at the terminal devices.

6.5.2 Transport layer

The transport layer conveys media from source sites to viewing sites. Considering security for content protection and latency of the network, this layer should be realized by next generation networks and IMT-2020 networks.

7 Candidate technologies for ILE functions

7.1 Ultra-realistic sound using wave field synthesis

One of the reconstructing technologies for the sound field is wave field synthesis (WFS). Loudspeaker arrays for WFS for surround sound systems are a way to enhance reality. WFS aims at reproducing an arbitrary sound field by using secondary sources. Typical implementations of WFS use linear distribution of secondary sources for reproducing a planar sound field.

7.2 Video stitching for surround view with super high definition

In order to achieve a high-realistic sensation with ultra-wide video, the ordinary 16:9 television video is not suitable for ILE services, thus, new algorithms and system architecture are required to capture video using several 4K cameras arranged horizontally. Figure 3 shows an example of a real time composition system using five 4K cameras. To process the large amount of 4K image data at high speed in real time, the data are partitioned and a mechanism is used that enables successive frames to be processed without waiting for positioning from the previous frame, which is necessary to suppress flicker. These innovations make real-time (4K60p) processing possible.

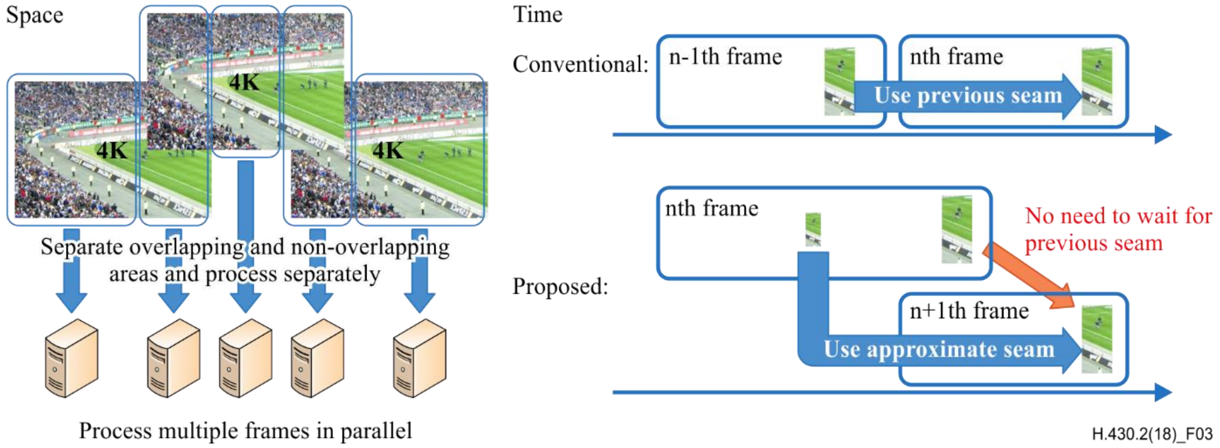

Figure 3 – Ultra-wide image composition

7.3 Media transport functions

One of the key features on ILE services is the transport of multiple live streaming content such as video and audio with spatial data stream of objects and other stage effects information. MMT [b-ISO/IEC 23008-1] can be utilized for this stream information for transmitting these media information synchronously.

7.4 Arbitrary background real-time object extraction function

In order to reconstruct high-realistic images such as pseudo 3D at the viewing sites, objects such as athletes or performers need to be extracted in real-time from the video captured at the event sites. For this purpose, a special photographic technique called chroma-keying is usually used. For chroma-keying, a screen in a single colour such as blue or green is placed behind the object so that the difference in colour between the object and the background screen can be used to extract the object. However, this technique cannot be used for actual events such as sports games and music concerts because backgrounds are not single colour. ILE services need accurate arbitrary background real time object extraction from video taken of the space surrounding the object, even from a sports field or stage.

8 General role model for providing ILE services

This clause describes the general role model for providing ILE services. Figure 4 shows the general role model based on high-level architecture with some interfaces. The model has four main roles:

– Production which may be content providers or content makers,

– Media processing for ILE services,

– Media transport, and

– Presentation at viewing sites.

There are three interfaces: IF-1 is between production and media transport, IF-2 is between media transport and presentation and IF-3 is between media transport and media processing.

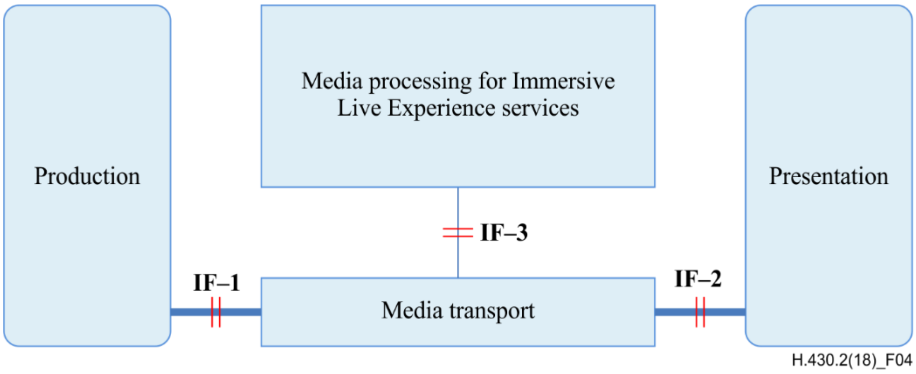

Figure 4 – General role model for providing ILE services

IF-1

This interface is defined as source content information which includes video, audio, lighting, and spatial information streaming.

IF-2

This interface is defined as asset information designed for each viewing site.

IF-3

This interface is defined as media processing information which includes source content information and asset information.

Appendix I

Examples of candidate technologies for functions of ILE

(This appendix does not form an integral part of this Recommendation.)

I.1 Implementation of WFS

Figure I.1 shows a block diagram of one implementation of WFS for an ILE system [b-Tsutsumi]. Figure I.2 shows an implementation of a front array for audio spatialization installed at the bottom of ta stage (the box surrounded by the dotted line) and a rear loudspeaker array for audio spatialization special effects is shown in Figure I.3. In this example, there are eight loudspeakers and one sub-woofer for surround sound reproduction. All the loudspeakers were connected to each other with Ethernet in a daisy chain configuration and controlled by digital audio network through Ethernet (DANTE) protocol. However, audio signals were fed through XLR cables from interfaces (Andiamo 2 XT by DirectOut Technologies).

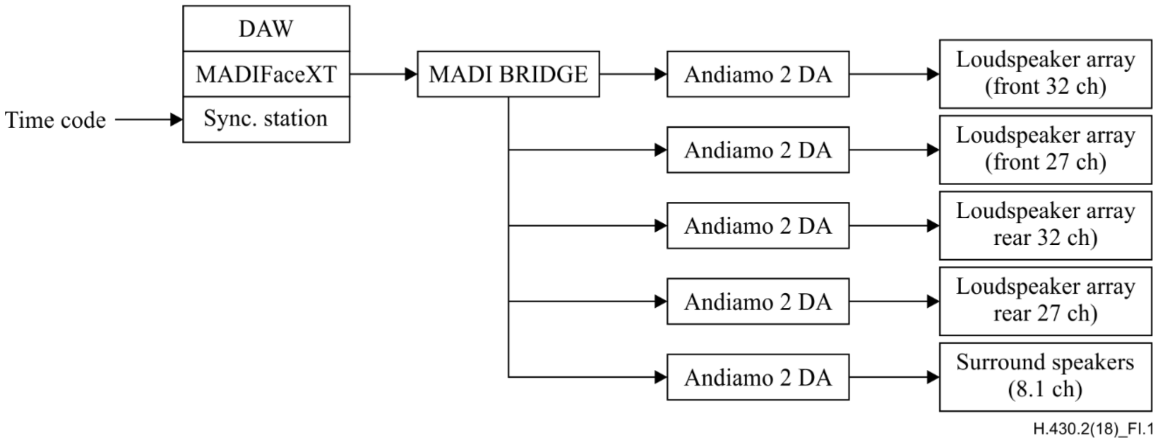

Figure I.1 – Block diagram of immersive sound reproduction system

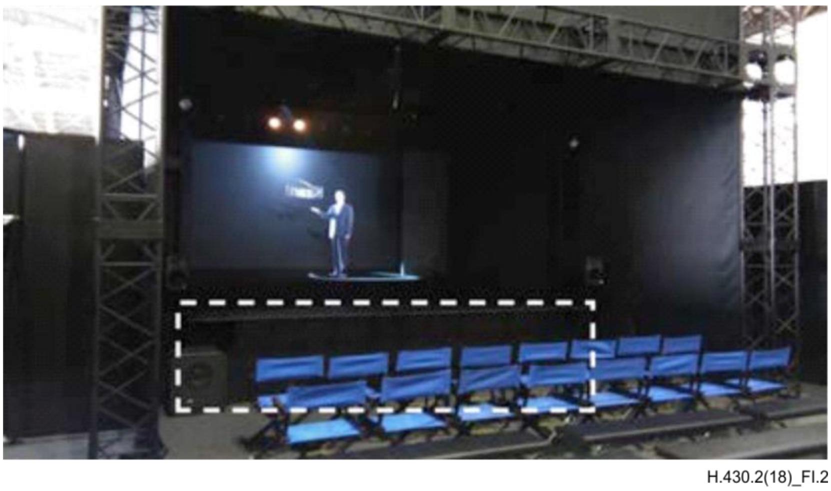

Figure I.2 – Front array embedded at the bottom of the stage (box with dotted line)


H.430.2(18)_FI.3

Figure I.3 – Rear loudspeaker array

A linear distribution of secondary sources was used along the x-axis for reproducing sound fields of focused sources defined at (x, y) = (0.0, 1.0), emitting monochromatic signals of 1 kHz. Assuming sound velocity as 343 m/s, the distance between reference lines and the x-axis was set to 3 m and the distance between loudspeakers was set to 0.075 m. Fifty-nine loudspeakers along the x-axis were used as secondary sources.

Since the sound field was assumed to be 1.1 m in height, i.e. the height of the ears of the sitting audience members, all the secondary sources were arranged in one line at the height of the loudspeakers (0.9 m).

I.2 Implementation of super high-definition video stitching

Details of an implementation of super high-definition video stitching are shown in [b-Akutsu].

I.3 Implementation of media transport functions

In order to synchronize and transport video and audio combined with spatial information, MMT, an optimized protocol for media synchronization, can be utilized with definitions for MMT signalling to describe 3D information such as the size, position and direction of MMT assets, see Figure I.4. This technology makes it possible to correlate physical spatial parameters such as the size and position of the display device with asset data (frame pixel data) so that the space can be reconstructed with high realism at the destination at the intended size. In addition, transmission of the digital multiplex (DMX) signals commonly used in production to control stage lighting and audio devices together with the MMT assets enables realistic presentations that accurately synchronize remote stage equipment with the media [b-Akutsu].

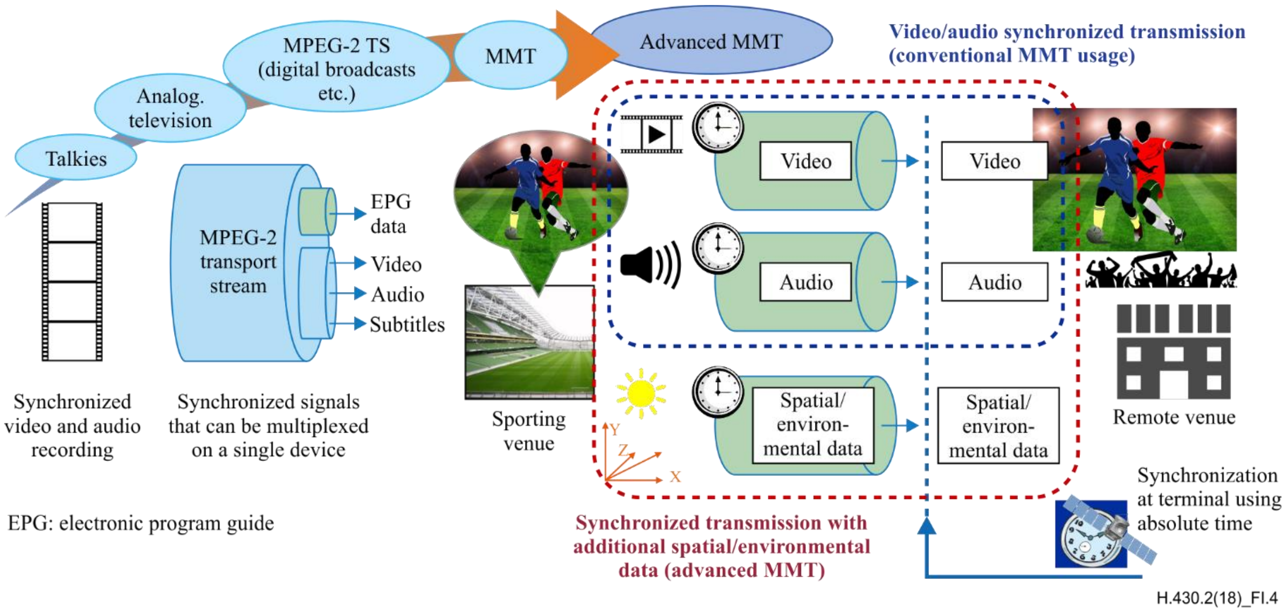

Figure I.4 – Media synchronization for realistic sensation (Advanced MMT)

I.4 Implementation of arbitrary background real-time object extraction function

An overview of the actual implementation of an arbitrary background object extraction process is shown in Figure I.5 [b-Nagata]. The object extraction uses sensors in addition to the imaging camera to capture the object. This enables constraints on object extraction to be eliminated so that the range of applications can be expanded as much as possible in the future.

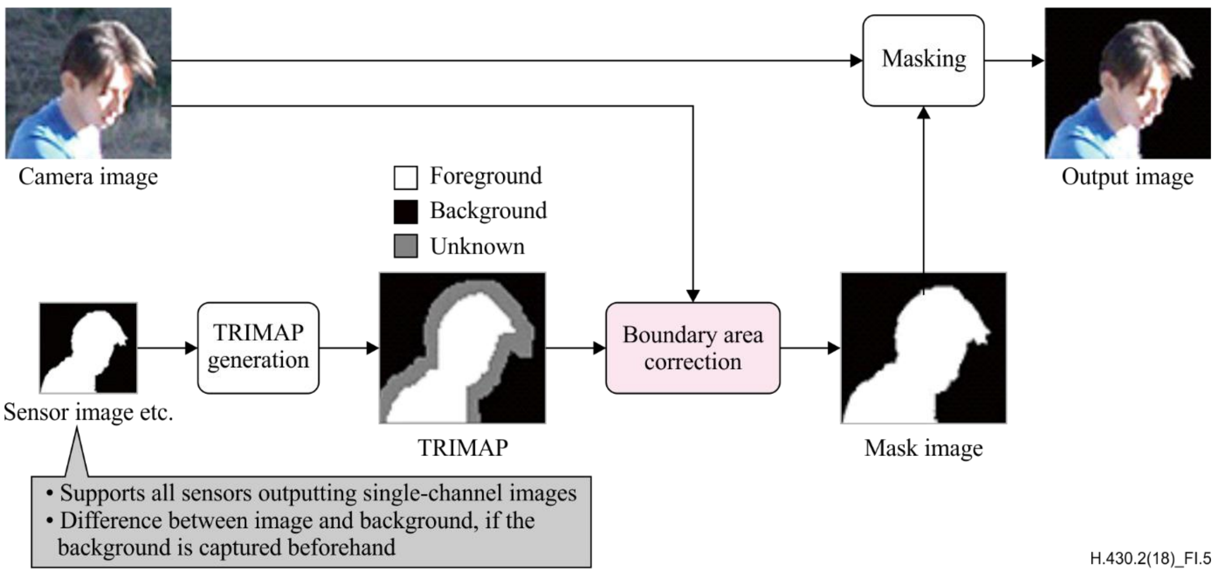

Figure I.5 – Overview of arbitrary background object extraction process

Processing can be divided broadly into two stages. In the first stage, a sensing device and the background image (only if it can be captured ahead of time) are used to roughly identify the object area. In this stage, data (a trimap) are generated and pixels are labelled into three categories: foreground, background and unknown. Then, pixels in the trimap in the region labelled 'unknown' (for which the foreground and background could not be distinguished during the rough identification stage and is assumed to contain the object boundary) are rapidly further classified as foreground and background and corrections are applied to the boundary.

In the second stage, some objects targeted for extraction have an extremely complex boundary, for example target objects are with untargeted objects in background. The boundaries of target objects must be represented in fine detail to avoid degrading the sense of realism. In such cases, pixels in the unknown region are classified as foreground or background, but they are also assigned a transparency value called an α (alpha) value, which provides a more natural-looking boundary.

Further work on the arbitrary background real-time object extraction technology is required to increase speed, robustness and detail.

Bibliography

[b-ITU-T H.430.1] Recommendation ITU-T H.430.1 (2018), Requirements for immersive live *experience (ILE) services*.

[b-ITU-T H.430.3] Recommendation ITU-T H.430.3 (2018), *Service scenario of immersive live experience (ILE)*.

[b-ISO/IEC 23008-1] ISO/IEC 28008-1:2014, *Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 1: MPEG media transport (MMT)*.

[b-Akutsu] Akihito Akutsu, Akira Ono, Hideaki Takada, Yoshihide Tonomura, and Maiko Imoto (2017), *2020 Public Viewing – Kirari! Immersive Telepresence Technology*, NTT Technical Review.

[b-Nagata] H. Nagata, H. Miyashita, H. Kakinuma, M. Yamaguchi (2017), *Real-time Extraction of Objects with Arbitrary Backgrounds*, NTT Technical Review, Vol. 15, No. 12. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201712fa7_s.html>

[b-Tsutsumi] Kimitaka Tsutsumi, Hideaki Takada, and Akira Ono (2017), *Immersive Live Experience Using Portable Loudspeaker Arrays for Wave Field Synthesis in Exhibitions*, Audio Engineering Society.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	**Transmission systems and media, digital systems and networks**
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound-programme and television-transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2019