Immersive visual media – MPEG-I: 360 video, virtual navigation and beyond

Marek Domański, Olgierd Stankiewicz, Krzysztof Wegner, Tomasz Grajek
Chair of Multimedia Telecommunications and Microelectronics,
Poznań University of Technology, Poland
{kwegner, ostank}@multimedia.edu.pl

Invited paper

Abstract **– In this paper we consider immersive visual media that are currently researched within scientific community. These include the well-established technologies, like 360-degree panoramas, as well as those being intensively developed, like free viewpoint video and point-cloud based systems. By the use of characteristic examples, we define the features of the immersive visual media that distinguish them from the classical 2D video. Also, we present the representation technologies that are currently considered by the scientific community, especially in the context of standardization of the immersive visual media in the MPEG-I project recently launched by ISO/IEC.**

Keywords – Immersive video; free viewpoint; 3D 360 video; virtual reality; MPEG-I; Future Video Coding

I. INTRODUCTION

The word *immersive* comes from Latin verb *immergere*, which means to dip, or to plunge into something. In the case of digital media, it is a term used to describe the ability of a technical system to absorb totally a customer into an audiovisual scene. *Immersive multimedia* [3] may be related to both natural and computer-generated content. Here, we are going to focus on the natural content that originates from video cameras, microphones, and possibly is augmented by data from supplementary sensors, like depth cameras. Such content is sometimes described as *high-realistic* or *ultra-realistic*.

Obviously, such natural content usually needs computer preprocessing before being presented to humans. A good example of such *interactive* content is spatial video accompanied by spatial audio that allows a human to virtually walk through a tropical jungle that is full of animals that are not always visitor-friendly. During the virtual walk, a walker does not scare the animals and may choose a virtual trajectory of a walk, may choose the current view direction, may stop and look around, hear the sounds of jungle etc. The respective content is acquired with the use of clusters of video cameras and microphones, and after acquisition must be preprocessed in order to estimate the entire representation of the audiovisual scene. Presentation of such content mostly needs rendering, e.g. in order to produce video and audio that corresponds to a specific location and viewing direction currently chosen by a virtual jungle explorer. Therefore, presentation of such content may also be classified as presentation of *virtual reality* although all the content represents real-world objects in their real locations with true motions (see e.g. [1]).

Obviously, the immersive multimedia systems may be also aimed at the computer-generated content, both standalone or mixed with natural content. In the latter case, we may speak about *augmented reality* that is related to "a computer-generated overlay of content on the real world, but that content is not anchored to or part of it" [1]. Another variant is *mixed reality* that is "an overlay of synthetic content on the real world that is anchored to and interacts with the real world contents". "The key characteristic of mixed reality is that the synthetic content and the real-world content are able to react to each other in real time" [1].

The natural immersive content is produced, processed and consumed in the path depicted in Fig. 1. As shown in Fig. 1, the immersive multimedia systems usually include communication between remote sites. Therefore such systems are also referred as *tele-immersive*, i.e. they serve for *highly realistic sensations communication* (e.g. [2]).

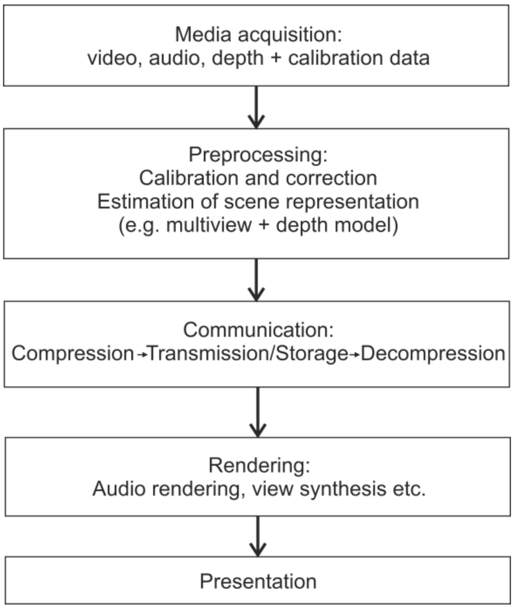
Media acquisition:
video, audio, depth + calibration data
Preprocessing:
Calibration and correction
Estimation of scene representation
(e.g. multiview + depth model)
Communication:
Compression→Transmission/Storage→Decompression
Rendering:
Audio rendering, view synthesis etc.
Presentation

Fig. 1. The processing path of immersive media.

The first block of the diagram from Fig. 1 represents *acquisition* of data that allows reconstruction of a portion of an *acoustic wave field* [4] and a *lightfield* [5], respectively. The audio and video acquisition using a single microphone and a single video camera is equivalent to the acquisition of a single spatial sample from an acoustic wave field and a lightfield, respectively. Therefore, the immersive media acquisition means acquisition of many spatial samples from these fields that would allow reconstruction of substantial portions of these fields. Unfortunately, such media acquisition results in huge amount of data that must be processed, compressed, transmitted, rendered and displayed.

Obviously, for immersive audio systems, the problems related to large data volume are less critical. Moreover, from the point of view of the necessary data volume, the human auditory system is also less demanding than the human visual system. These are probably the reasons, why the immersive audio technology seems to be currently more mature than the immersive video technology. There exist several spatial audio technologies like *multichannel audio* (starting from the classic 5.1 and going up to the forthcoming 22.2 system), *spatial acoustic objects* and *higher order ambisonics* [6]. During the last decade, the respective spatial audio representation and compression technologies have been developed and standardized in MPEG-D [7], [8] and MPEG-H Part 3 [9] international standards. The spatial audio compression technology is based on coding of one or more stereophonic audio signals and additional spatial parameters. In that way, this spatial audio compression technology is transparent for the general stereophonic audio compression. Currently, the state-of-the-art audio compression technology is *USAC (Unified Speech and Audio Coding)* standardized as MPEG-D Part 3 [10] and *3D audio* standardized as MPEG-H Part 3 [9].

Also the presentation technology has been well advanced for spatial audio. These developments are not only related to the systems with high numbers of loudspeakers but also to binaural rendering for headphone playback using *binaural room impulse responses (BRIRs)* and *head-related impulse responses (HRIRs)* that is a valid way of representing and conveying an immersive spatial audio scene to a listener [11].

The above remarks conclude the considerations related to immersive audio in this paper that is focused on immersive visual media. For the immersive video, the development is more difficult, nevertheless the research on immersive visual media is booming recently.

Considering the immersive video, one has to mention 360-degree video that is currently under extensive market deployment. The 360-degree video allows at least to watch video in all directions around a certain position of a viewer. On the other hand, other systems, like *free-viewpoint television* [20] or virtual navigation, allow user to freely change location of viewpoint. The most advanced systems, called omnidirectional 6DoF [23], extends 360 degree video and free-navigation, in order to allow the user to both look in any direction and virtually walk thought prerecorded world.

These interactive services provide for a viewer an ability to virtually walk around a scene and watch a dynamic scene from any location on the trajectory of this virtual walk [21], [22]. Therefore, in popular understanding, the 360-degree video is treated as a synonym to the immersive video, e.g. see Wikipedia [19].

Among many issues, the technical progress in immersive video is inhibited by the lack of satisfactory compression technology and by the lack of efficient displays producing high-realistic spatial sensations.

II. Immersive Video Compression

Currently, about 70% of all Internet traffic is Internet video traffic [12] that almost exclusively corresponds to monoscopic single-view video. Therefore, an upgrade of a substantial portion of the video traffic to the immersive video is undoable using the existing technology, because it would result in drastic increase of the demand for bandwidth in the global telecommunication network. The progress must be done both in single-view video compression as well as in spatial video compression that usually exploits the existing general video compression technology.

A. General video compression technology

In the last two decades, consecutive video coding technologies, i.e. MPEG-2 [13], AVC – Advanced Video Coding [14], and HEVC – High Efficiency Video Coding [15] have been developed thanks to huge research efforts. For example, the development and the optimization of HEVC needed an effort that may be measured in thousands of man-years.

When considering the three abovementioned representative video coding standards, some regularity is visible [16]. For each next generation, for a given quality level, the bitrate is halved. The temporal interval of about 9 years occurs between the consecutive technology generations of video coding. During each 9 years cycle the available computational power is increased by a factor of about 20-25, according to the Moore law. This computational power increase may be consumed by the next generation of more sophisticated video encoders.

For television services, for demanding monoscopic content, the average bitrate B may be very roughly estimated by the formula [16, 17, 18]

$$B \approx A \cdot V \;\; [\mathrm{Mbps}] \,, \qquad (1)$$

where: A is the technology factor: A=1 for HEVC, A=2 for AVC, A=4 for MPEG-2, and V is the video format factor,

- V=1 for SD – *Standard Definition* (720×576, 25*i*),
- V=4 for HD – *High Definition* (1920×1080, 25*i*),
- V=16 for UHD – *Ultra High Definition* (3840×2160, 50*p*).

Interestingly, there is already some evidence that this prediction will be true for HEVC successor, tentatively called FVC (Future Video Coding). FVC technology is currently being developed by joint effort of ISO/IEC MPEG (International Organization for Standardization / International Electrotechnical Commission, Motion Picture Experts Group) and ITU VCEG (International Telecommunication Union, Video Coding Experts Group) that formed JVET (Joint Video Exploration Team). The recent results demonstrate about 30% bitrate reduction using FVC over HEVC. The finalization of FVC technology and the inclusion of its specification in the forthcoming MPEG-I standard is expected around years 2020-2021. This would roughly match the abovementioned prediction based on the 9-year cycle. The goal for this new video compression technology is to provide bitrate reduction of about 50% as compared to the state-of-the-art HEVC technology. Therefore, for this new video compression technology the technology factor will be $A = 0.5$ in (1) [18].

The abovementioned video compression technologies are related to particular applications:

- MPEG-2 has enabled development of the standard-definition digital television (SDTV),
- AVC is widely used to high-definition services also in internet,
- HEVC has been developed for ultra-high definition services.

Unfortunately, the replacement of one of the abovementioned applications by the next higher-level application increases the required bitrate by a factor of about 4, while the next generation of video compression technology reduces the bitrate by a factor of about 2 only (cf. Fig. 2). The increase of the bitrate due to introduction of immersive video is expected to be significant.

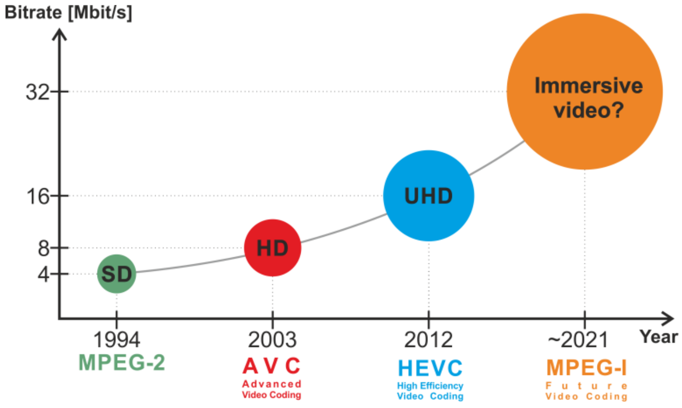

Fig. 2. Bitrates for major applications of the consecutive video compression generations.

Unfortunately, even further increase of the requested bitrates will be necessary in order to accommodate the introduction of the High Dynamic Range and Wide Color Gamut [24] into the Ultra High Definition video.

B. Video compression for spatial and immersive video

For a natural 3D visual scene modeling, several models are considered in the literature: object-based [25,26], ray space [20,27], point cloud-based [29], and multiview plus depth (MVD) [28]. The compression of the first types of representations is yet under development, while the MVD representation has been successfully used and standardized on basis of AVC [14] and HEVC [15] technologies. Currently, further standardization of MVD compression is also considered [30, 31].

The AVC [14, 33] and HEVC [15, 34] standards provide the profiles for the *multiview coding* as well as for the *3D video coding*.

The main idea of the *multiview coding* is to compress video acquired using several synchronized cameras, and to exploit the similarities between neighboring views. One view is encoded like a monoscopic video, i.e. using the standard intraframe and temporal interframe predictions. The produced bitstream constitutes the base layer of the multiview video representation. For the other views, in addition to the intraframe and interframe predictions, the inter-view prediction with disparity compensation may be used. In such prediction, a block of samples is predicted using a reference block of samples from a frame from another view in the same time instant. The location of this reference block is pointed out by the disparity vector. This inter-view prediction is dual to the interframe prediction, but the motion vectors are replaced by the disparity vectors, and the temporal reference frames are replaced by the reference frames from other views.

The extensions of the AVC and HEVC standards provide also the ability to encode the depth maps, where nonlinear depth representations are allowed [32].

The multiview coding provides the bitrate reduction of order 20-35%, sometimes reaching even 40% as compared to the simulcast coding [70]. These high bitrate reductions are achievable for video that is obtained from cameras densely located on a line, and then rectified in order to virtually set all the optical axes parallel and on the same plane. For sparse and arbitrary camera locations, the gain with respect to the simulcast coding reduces significantly. Therefore, the basic multiview video coding has nearly no importance for future compression of the immersive video.

Another option for considerations is 3D video coding**.** The distinction between multiview video coding and 3D video coding is not precise. The latter refers to: compression of the multiview plus depth representations and application of more sophisticated compression techniques of inter-view prediction. Great diversity of 3D video coding tools has been already proposed including prediction based on: view synthesis, inter-view prediction by 3D mapping defined by depth, advanced inpainting, coding of disoccluded regions, depth coding using platelets and wedgelets etc. [35, 36, 37, 38, 39, 40]. Some of these tools have been already included into the standards of 3D video coding: 3D High Profile of AVC [14, 41] and 3D Main Profile of HEVC [15, 34]. The latter defines the state-of-the-art technology for compression of multiview video with accompanying depth.

The 3D extension of HEVC is called 3D-HEVC. Similarly as in multiview coding in AVC, the standardization requirement was to reuse the monoscopic decoding cores for implementations. The multiview, 3D, and the scalable extensions of HEVC share nearly the same high-level syntax of the bitstreams. Therefore, for the standard, it was decided that view (video) encoding should not depend on the corresponding depth. Moreover, the 3D-HEVC provides additional prediction types that are not used in multiview coding:

1) Combined temporal and inter-view prediction of views that refers to pictures from another view and another time instant;
2) View prediction that refers to a depth map corresponding to the previously encoded view;
3) Prediction of depth maps using the respective view or a depth map corresponding to another view.

The compression gain of 3D-HEVC over the multiview profiles of HEVC is expressed by 12-23% bitrate reduction [71]. These compression gains are smaller when cameras are not aligned on a line. For circular camera arrangements, in particular with the angles between the camera axes exceeding 10 degrees, the gain over the simulcast coding falls below 15%, often being around 5%. In particular, for cameras sparsely located on an arc, the compression gains of 3D-HEVC over the simulcast HEVC are dramatically small. This observation stimulated research on the extensions of 3D-HEVC that use true 3D mapping for more efficient inter-view prediction [42, 43]. Such extension of 3D-HEVC has been proposed in the context of transmission of the multiview plus depth representations of the dynamic scenes in the future free-viewpoint television and virtual navigation systems [44]. Nevertheless, the results are not satisfactory yet, thus further progress is required. The state-of-the-art technology for MVD compression is definitely insufficient for such immersive video scenarios as virtual navigation.

3D video coding is currently a research topic for several groups around the world, and also future standardization activities are expected. In 2015, the MPEG-FTV, the body that was working within MPEG, was exploring possible 3D-HEVC extensions for efficient coding of multiview video taken from arbitrary camera positions. Probably, the next developments will be done not as extensions of HEVC but rather on the top of the forthcoming new general video coding technology developed as FVC (Future Video Coding) that will be probably standardized in the framework of the MPEG-I project. The expected gains for 3D video coding should come both from the more efficient general video coding, and from better tools of 3D video coding.

Instead of transmitting multiple views of a scene, one can represent the scene as a 3D cloud of colored points, i.e. as a *point cloud*. One can use such a point-based model instead of the MVD (multiview plus depth). The points in the model may be arranged as a raster of volumetric elements (*voxels*). Only the voxels on object boundaries need to be coded and transmitted. In a receiver, the 3D scene is rendered from the decoded points. A compression technology for the point clouds is still less advanced and less developed than that for MVD models, but the research on point cloud compression has accelerated recently [45]. Moreover, MPEG has also launched a standardization project on point cloud compression, thus further stimulating the research on this topic.

Although, substantial progress in the immersive video compression is needed, the recent developments in the adaptive streaming (like Dynamic Adaptive Streaming over HTTP – DASH [46, 47]) and media transportation (like MPEG Media Transport – MMT [48, 49]) provide well-advanced platforms for the immersive media delivery.

III. Displays

Apart from the video compression technology, the display technology is also not mature enough for the immersive video and images. Nevertheless, the situation is unequal for various display application areas.

A. Large displays and projection systems

One of the immersive display already in used for over a quarter of century [50] is a media cave. The caves are special rooms in which the images are projected on walls and ceiling. A variant of this approach is TiME Lab of Fraunhofer HHI with huge cylindrical display well on one side of a room [51]. Also display wells, or walls with background projection, are used around a viewer in order to produce the impression of immersion. Such solutions have been proposed also for the immersive telecommunication [52].

For signage applications large autostereoscopic projection systems (called super-multiview displays) are used. Such display systems provide viewers with high quality of experience of spatial sensation. Such systems [53] may display more than 60 million of pixels simultaneously in order to produce large number of views that correspond to potential gaze locations. Large number of views is needed for seamless parallax.

An extreme version of a glassless 3D display is the display installed by Japanese National Research and Development Agency (NICT) in a commercial center in Osaka [54, 55]. This back-projection display benefits from about 200 HD video projectors fed by a large cluster of computers that render the respective video streams. The size, the cost and the power consumption withdraw such 3D display system from a wider use.

B. Consumer displays

On the consumer 3D displays market, the displays that require glasses to properly separate the left and right view for the respective eyes are the most popular. Such glasses are not widely acceptable, and this is considered the main reason for the recent decline in 3D video.

The remedy is seen in autostereoscopic or lightfield displays that use the well-known liquid crystal (LCD) technology. In such displays the LCD panel displays a number of views that correspond to various potential gaze locations. A pattern of lenses on the display is used in order to direct the beams from individual views to the respective potential gaze locations. Multiple users can view difference view of the scene but simultaneously the number of views that are displayed grows, the sweet point becomes longer and the motion parallax tends to become seamless. Unfortunately, the view resolutions decrease for a given LCD panel resolution. The quality of experience becomes really good for displays with about 100 views that need 4K or 8K displays [56]. Nevertheless, such display suffer from high weight as they need many built-in components for rendering of the high number of parallel video streams. Also, their current cost is preventive for consumer entertainment applications.

Other 3D display types, like holographic displays are even less mature.

C. Head-mounted devices (HMDs)

The most advanced and recognized for immersive media display are head-mounted devices (HMDs) [57,58] that are blooming in all virtual-reality applications. There are two groups of such devices:

- simple and cheap devices where a smartphone is used as both a display and a processing unit,
- more advanced and expensive devices, like helmets, that are equipped with special single-eye displays.

Unfortunately, most of the HMDs currently available on the market lack high resolution needed for immersive experience. Moreover, the humans are extremely sensitive to the delays between head motion and the image displayed in HMD. Therefore, extremely low latency is required which is challenging for local systems and nearly killing factor for all network-related systems.

Therefore, we conclude that the display technology is not yet mature enough in order to accommodate the needs of future immersive video systems.

IV. Immersive Visual Technologies

In order to absorb the user into immersive reality, the underlying technology has to convince our senses. The most basic is to present vision to the eyes of the viewer. It is however not enough to fool our brain entirely. The level of immersion can be increased if the following features are addressed:

- *Rotation*. The ability to look around freely with 3 degrees of freedom (DoF), e.g. to yaw, pitch or roll, allows human brain to construct holistic model of the environment. This process is crucial to provide true immersive experience. Therefore, the views presented to the user's eyes should follow rotation of the head.
- *Motion*. The ability of the user to move in all directions (3 DoF) improves the level of immersion in two ways. First, it enables motion parallax, which helps brain to perceive the depth and cope with occlusions. Second, it allows the user to explore.
- *Join rotation and motion***.** Both of these features together grant the user with 6 degrees of freedom. Thanks to the synergy, the user can witness the presented reality without bounds.
- *Latency***.** Our brains are very vulnerable to the differences in time of perception of information coming from different senses. The mismatch causes motion sickness, which can be avoided by minimizing overall latency of the system.
- *Binocular vision*. The human visual system employs information from both of eyes to sense the depth of the scene. Without the proper depth sensation, the scene is perceived as flat and unnatural.
- *Resolution*. For the example of HMDs, the displays are mounted very close to the user's eyes and therefore the amount of pixels must be sufficient in order to avoid aliasing. This is in particular important when the user moves or rotates very slightly, which results in unnatural jumps of edges in the perceived image by single pixel distance.
- *Self-embodiment***.** The ability to see parts of its own body convinces the user about being part of the presented reality.
- *Interactivity*. Allowing the user to manipulate objects provides strong premises about integrity of the presented reality.

In this section, we characterize existing and emerging technologies that provide immersive experience. Not all of them support all of the mentioned immersive features, which of course degrade the attained level of immersiveness.

A. Monoscopic 360 video

360-degree video conveys the view of a whole panorama seen from given point (Fig. 3). Practically, 360-degree video is most often captured by a set of at least 4 to 6 cameras looking outwards (Fig. 4). Images from individual cameras are then stitched [59,60] together in order to produce a single panorama view.

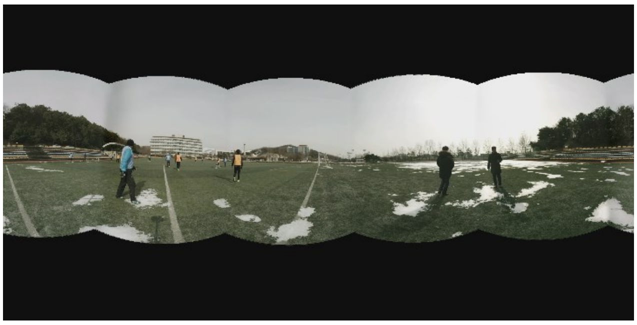

Fig. 3. A panorama represented in 360-degree video. Example frame from "Soccer360" sequence [67]. Image used thanks to the courtesy of Electronics and Telecommunications Research Institute (ETRI), Korea.

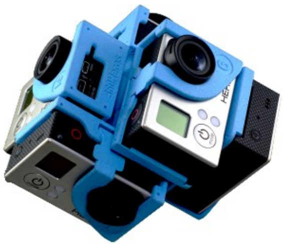

Fig. 4. 360-degree camera built from six GoPro Hero 3 action cameras (photo by the authors).

The data in 360-degree video is represented in a form resembling classical 2D image but with the pixel coordinates interpreted as values related to angles instead of positions on flat image plane of the camera. The distance between the left and the right edge is 360 degrees and thus the edges coincide. The particular mapping of longitude and latitude to pixel coordinates may be specified in various ways. The most commonly known are equirectangular [61] and cylindrical projection [62] (Fig. 5), both having advantages and disadvantages related to the represented range and resolution of angles. In ideal model of the acquisition, each column of pixels is captured by a separate outward-looking (at different longitude) camera with a very narrow horizontal field of view (FoV) and some vertical FoV. In the case of equirectangular projection, vertical FoV is 180 degrees and less in the case of cylindrical projection.

The variety of possible mappings is a challenge for standardization [63]. One of the currently considered solutions is to use mesh-based mapping instead of a set of selected mathematical formulations. The works in MPEG are still undergoing, but the specification document, named Omnidirectional Media Application Format (OMAF) [64] is expected to be finished by the end of 2017 [65].

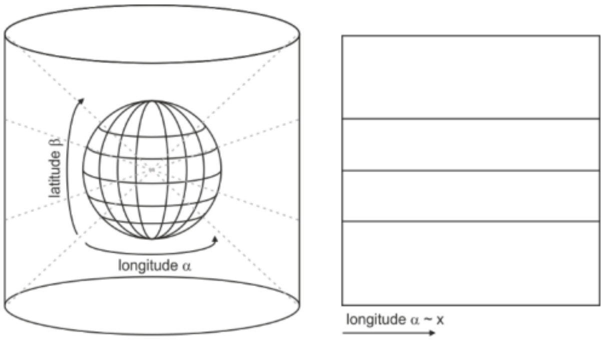

Fig. 5. A panorama represented in 360-degree video.

By presenting a selected fragment of 360-video, it is possible to allow the user to rotate (yaw, pitch or roll). It is not possible though to move. Therefore, 360-video is often said to be 3 DoF (degrees of freedom). Also, because both eyes are seeing the same panorama, there are no depth impression.

B. Stereoscopic 360 video

Stereoscopic 360-degree video is an extension of idea of 360-degree video, in which there are two panoramas of the scene – one for the left and one for the right view. Often it is referred to as "3D 360". The two panoramas are typically arranged in top/bottom or left/right manner in a single image (Fig. 6).

Just like in 360-degree video, each row of pixels is related to outward-looking camera with a very narrow horizontal FoV. The difference between the left and the right panorama is displacement from the center of the cameras (Fig. 7).

The usage of two panoramas allows presentation of different views for the left and for the right eye, which produces sensations of depth in the scene. Unfortunately, the depth sensations are limited to the regions near the equator of the projection because at the poles both images are the same (the cameras presented in Fig. 7 lay on the same horizontal plane).

Also, the user is not allowed to move, not even slightly, which yields unnatural 3D impression with head movements.

It is expected that delivery formats for 3D 360 video will be standardized by MPEG until the end of 2018 [66].

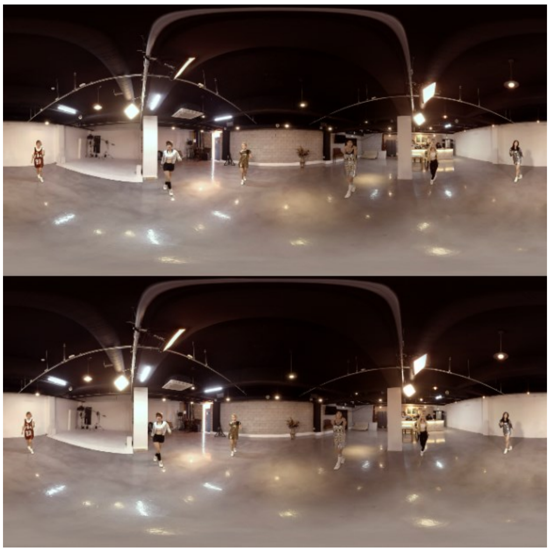

Fig. 6. Example of panorama for the left and for the right view in top/bottom stereoscopic 360-degree video from "Dancer360" test sequence [67]. Image used thanks to the courtesy of Electronics and Telecommunications Research Institute (ETRI) , Korea.

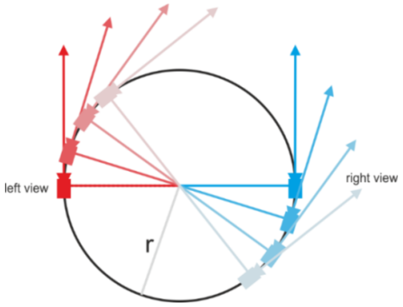

Fig. 7. Model of 3D 360 video capturing. Each row of pixels in panoramas is captures by a camera with very narrow horizontal field of view. The left and for the right view camera are displaced from the center.

C. Binocular 3D 360 video

The aim for binocular 3D 360-degree video technology is to overcome the biggest limitations of stereoscopic 360 video: limitation of depth sensation apart from the equator and the lack of motion parallax. Because the user will not be allowed to move freely, but only slightly, this technology is often to be referred to be 3 DoF+ (plus) as it does not provide full 6 DoF.

Rendering of views with motion parallax requires some depth information about the scene, e.g. further objects move slower in perspective than closer objects. Currently there are two solutions considered within MPEG. The first one is the usage of layered stereoscopic 360 video, where each layer is assigned a constant depth level. The second one is usage of depth maps, which convey individual depth information for each point in the image. Of course, both of those require information about the depth of the scene, which has to be acquired directly (e.g. by means of Time-of-Flight cameras) or estimated algorithmically. Already there are works that report techniques for depth estimation from stereoscopic 360 video. In paper [68] authors show that it is possible to approximate depth estimation from 3D 360 video with classical stereoscopic depth estimation. The conclusions are that the classical stereoscopic depth estimation formula (2), where Z is the sought distance, f is focal length of the model of cameras used in depth estimation, b is baseline distance between them and d is disparity between matched features (e.g. points):

$$Z = \frac{f \cdot b}{d} \quad , \qquad (2)$$

can be used for 360 video with a very small error with the following mathematical parameters (3), without knowing the physical parameters of the capturing camera rig:

$$f = \frac{W}{2\pi} \quad ; \quad b = 2 \cdot r \quad , \qquad (3)$$

where W is the width of the panorama (in pixels) and r is radius of the camera rig (the scale of the space). The results of such an approach are presented in Fig. 8.

Binocular 3D 360 video is expected to be standardized by MPEG until the end of 2019.

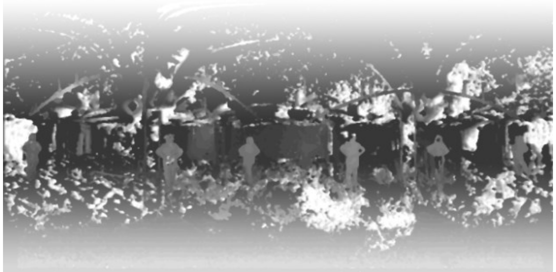

Fig. 8. Depth map estimated from 3D panorama (Fig. f4) with the use of approach from [68] for “Dancer360” test sequence [67].

D. Free viewpoint video

In a free viewpoint video system, the user is allowed to freely select the point of view: the direction of looking and the position. Each of such provides the user 3 degrees of freedom (Dof), and therefore, thanks to the synergy of both, such systems provide the user 6 DoF. The difference between binocular 3D 360 and free viewpoint is the video format. In spite of panoramic images, the most promising free viewpoint video systems use Multiview Video plus Depth (MVD) representation format. In MVD the scene is captured with a limited number of cameras (e.g. 10) positioned around the scene (Fig. 9). Each view captured by a camera is associated with corresponding depth map.

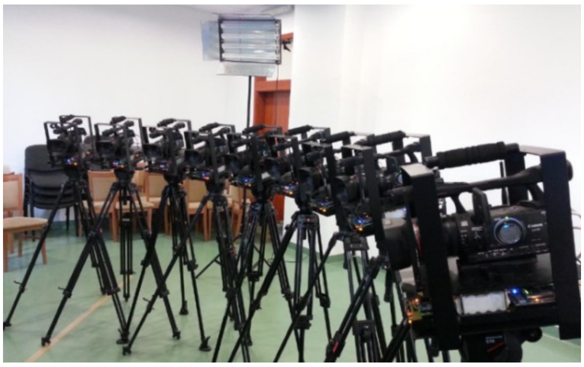

Fig. 9. Experimental free viewpoint acquisition system built at Poznan University of Technology.

The views and the depth maps together are used to render desired view to the user’s left and the right eye, e.g. with use of Depth-Image Based Rendering (DIBR) techniques. Of course, the viewing is limited to the regions which are captured with the cameras, and thus, the experience resembles watching of the scene through a clear window. Therefore, in MPEG, free viewpoint video systems are referred to as “windowed 6DoF” (Fig. 10). This also limits the freedom of the user in practical applications. For example, if the cameras are looking outwards, then, similarly to Binocular 3D 360 video, the allowed motion of the user is very small. On the other hand, if cameras are on a side of the scene, then the motion of the user can move almost without bounds, but the allowed rotation is very limited.

All technical aspects of free viewpoint video systems are considered to be very difficult research problems and are currently subjects of extensive research. Only few experimental systems provide satisfactory quality of experience. An example of such is the experimental Free Viewpoint Television system developed at Poznan University of Technology (Fig. 9). The developed algorithms used within the system allow for estimation of high quality depth maps that can be used to synthesize virtual views to the user (Fig. 11). Although for some of the test sequences the attained quality is satisfactory, the works to improve the results are still in progress. Due to challenges in this field, technologies related to free viewpoint video are not expected to be standardized before 2021.

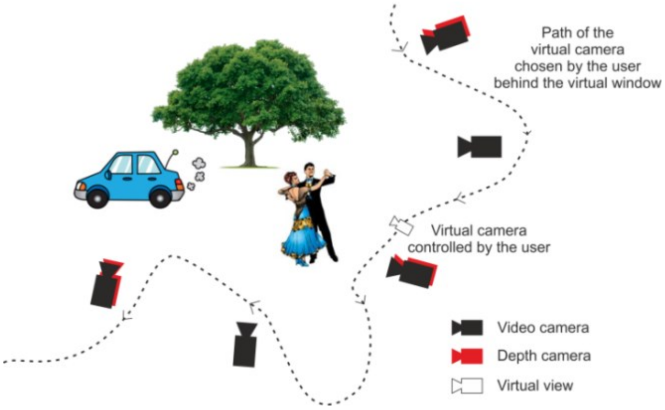
Path of the virtual camera chosen by the user behind the virtual window
Virtual camera controlled by the user
Video camera
Depth camera
Virtual view

Fig. 10. Illustration of free navigation in windowed 6 degrees of freedom (6DoF) scenario.

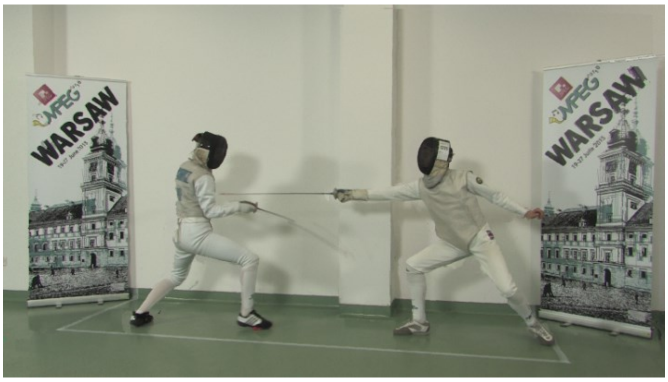

Fig. 11. Example of synthesized views of “Poznan Fencing” [21,69] sequence generated with experimental free viewpoint system built at Poznan University of Technology.

V. Conclussions

We have presented various immersive visual media, including 360-degree panorama video, stereoscopic (3D) 360 video, binocular 360 video, free viewpoint video and point-cloud representation. For each of presented examples, we have considered the supported features which determine the attained level of immersiveness. As shown, this level is different among the considered technologies, and varies with technological complexity. In some of the cases, the technology is anticipated to be available in the not so close future. This fact is one of the motivations of works in ISO/IEC MPEG group on MPEG-I project, which aims at standardization of immersive visual media in phases. Due to the current plans, the first stage of MPEG-I, phase 1a, will target the most urgent market needs, which is specification of 360 video projection formats – Omnidirectional Media Application Format (OMAF) [64,65]. The next phase of MPEG-I, 1b [66], will the extend specification provided in 1a towards 3 DoF+ applications. The phase 2, which is intended to start from about 2019, will aim at addressing 6 DoF applications like free viewpoint video. Therefore, it can be summarized that the technologies that are already settles will be standardized first and will be followed by extensions related to technologies that will mature later.

VI. Acknnowledgement

This work has been supported by the public funds as a DS research project 08/84/DSPB/0190.

References

[1] EBU Technical Report TR 039, “Opportunities and challenges for public service media in vr, ar and mr”, Geneva, April 2017.

[2] T. Ishida, Y. Shibata, “Proposal of tele-immersion system by the fusion of virtual space and real space”, 2010 13th International Conference on Network-Based Information Systems (NBiS), Takayama, Gifu, Japan, 2010.

[3] F. Isgro, E. Trucco, P. Kauff, O. Schreer, “Three-dimensional image processing in the future of immersive media”, IEEE Trans Circuits Syst. Video Techn., vol. 14, 2004, pp. 288 – 303.

[4] J. Benesty, J. Chen, and Y. Huang, “Microphone array signal processing”, Springer-Verlag, Berlin, 2008.

[5] M. Ziegler, F. Zilly, P. Schaefer, J. Keinert, M. Schöberl, S. Foessel, “Dense lightfield reconstruction from multi aperture cameras”, 2014 IEEE Internat. Conf. Image Processing (ICIP), Paris 2014, pp. 1937 – 1941.

[6] J. Herre, J. Hilpert, A. Kuntz, .J. Plogsties, MPEG-H 3D Audio—The new standard for coding of immersive spatial audio , IEEE Journal of Selected Topics In Signal Processing, vol. 9, 2015, pp.770-779.

[7] ISO/IEC IS 23003-1: 2007, “MPEG audio technologies -- Part 1: MPEG Surround”.

[8] ISO/IEC IS 23003-2: 2016 (2nd Ed.) “MPEG audio technologies -- Part 2: Spatial Audio Object Coding (SAOC)”.

[9] ISO/IEC IS 23008-3: 2015, “High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio”.

[10] ISO/IEC IS 23003-2: 2016 (2nd Ed.) “MPEG audio technologies -- Part 3: Unified Speech And Audio Coding (USAC)”.

[11] J. Blauert, Ed., “Technology of binaural listening”, Springer-Verlag, Berlin/Heidelberg, 2013.

[12] Cisco, “Visual Networking Index: Forecast and Methodology, 2015–2020”, updated June 1, 2016, Doc. 1465272001663118.

[13] ISO/IEC IS 13818-2: 2013 and ITU-T Rec. H.262 (V3.1) (2012), “Generic coding of moving pictures and associated audio information – Part 2: Video”.

[14] ISO/IEC IS 14496-10: 2014 “Coding of audio-visual objects - Part 10: Advanced Video Coding”and ITU-T Rec. H.264 (V9) (2014), “Advanced video coding for generic audiovisual services” .

[15] ISO/IEC Int. Standard 23008-2: 2015 “High efficiency coding and media delivery in heterogeneous environment – Part 2: High efficiency video coding”and ITU-T Rec. H.265 (V3) (2015), „High efficiency video coding”.

[16] M. Domański, T. Grajek, D. Karwowski, J. Konieczny, M. Kurc, A. Łuczak, R. Ratajczak, J. Siast, J. Stankowski, K. Wegner, “Coding of multiple video+depth using HEVC technology and reduced representations of side views and depth maps,” 29th Picture Coding Symposium, PCS, Kraków, May 2012.

[17] M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, Ł. Kowalski, M. Kurc, A. Łuczak, D. Mieloch, R. Ratajczak, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, “Methods of high efficiency compression for transmission of spatial representation of motion scenes”, IEEE Int. Conf. Multimedia and Expo, Torino 2015.

[18] M. Domański, “Approximate video bitrate estimation for television services”, ISO/IEC JTC1/SC29/WG11 MPEG2015, M36571, Warsaw, June 2015.

[19] https://en.wikipedia.org/wiki/360-degree_video , as April 29th, 2017.

[20] M. Tanimoto, M. P. Tehrani, T. Fujii, T. Yendo “FTV for 3-D spatial communication”, Proc. IEEE, vol. 100, no. 4, pp. 905-917, 2012.

[21] M. Domański, M. Bartkowiak, A. Dziembowski, T. Grajek, A. Grzelka, A. Łuczak, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, “New results in free-viewpoint television systems for horizontal virtual navigation”, 2016 IEEE International Conference on Multimedia and Expo ICME 2016, Seattle, USA, July 2016

[22] G. Lafruit, M. Domański, K. Wegner, T. Grajek, T. Senoh, J. Jung, P. Kovacs, P. Goorts, L. Jorissen, A. Munteanu, B. Ceulemans, P. Carballeira, S. Garcia, M. Tanimoto, “New visual coding exploration in MPEG: Super-multiview and free navigation in free viewpoint TV”, IST Electronic Imaging, Stereoscopic Displays and Applications XXVII, San Francisco 2016.

[23] M.-L. Champel, R. Koenen, G. Lafruit, M. Budagavi “Working Draft 0.2 of TR: Technical report on architectures for immersive media”, ISO/IEC JTC1/SC29/WG11, Doc. MPEG-2017 N16918, Hobart, April 2017.

[24] ITU-R Rec. BT.2020-1 “Parameter values for ultra-high definition television systems for production and international programme exchange”, 2014.

[25] G. Miller, J. Starck, A. Hilton, “Projective surface refinement for free-viewpoint video,” 3rd European Conf. Visual Media Production, CVMP 2006, pp.153-162.

[26] A. Smolic, et al., “3D video objects for interactive applications.” European Signal Proc. Conf. EUSIPCO 2005.

[27] M. Tanimoto, “Overview of free viewpoint television”, Signal Processing: Image Communication, vol. 21, 2006, pp. 454-461.

[28] K. Müller, P. Merkle, T. Wiegand, “3D video representation using depth maps”, Proc. IEEE, vol. 99, pp. 643–656, April 2011.

[29] K.-Ch. Wei, Y.-L. Huang, S.-Y. Chien, “Point-based model construction for free-viewpoint tv,” IEEE Int. Conf. Consumer Electronics ICCE 2013, Berlin, pp.220-221.

[30] M. Tanimoto, T. Senoh, S. Naito, S. Shimizu, H. Horimai, M.Domański, A. Vetro, M. Preda, K. Mueller, “Proposal on a new activity for the third phase of FTV”, ISO/IEC JTC1/SC29/WG11 Doc. MPEG-2015 M30232, Vienna, July 2013.

[31] M. Domański, A. Dziembowski, K. Klimaszewski, A. Łuczak, D. Mieloch, O. Stankiewicz, K. Wegner, “Comments on further standardization for free-viewpoint television,” ISO/IEC JTC1/SC29/WG11 Doc.MPEG-2015 M35842. Geneva, October 2015.

[32] O. Stankiewicz, K. Wegner, M. Domański, "Nonlinear depth representation for 3D video coding", IEEE International Conference on Image Processing ICIP 2013, Melbourne, Australia, 15-18 September 2013, pp. 1752-1756.

[33] A. Vetro, T. Wiegand, G. J. Sullivan, “Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard”, Proceedings of the IEEE, vol. 99, 2011, pp. 626-642.

[34] G. Tech, Y. Chen, K. Müller J.-R. Ohm, A. Vetro, Y.-K. Wang, “Overview of the multiview and 3D extensions of high efficiency video coding”, IEEE Transactions on Circuits and Systems for Vieo Technology, vol. 26, No. 1, January 2016, pp. 35-49.

[35] Y. Chen, X. Zhao, L. Zhang, J. Kang, “Multiview and 3D video compression using neighboring block based disparity vector”, IEEE Transactions on Multimedia, Volume: 18, pp. 576 – 589, 2016.

[36] M. Domański, O. Stankiewicz, K. Wegner, M. Kurc, J. Konieczny, J. Siast, J. Stankowski, R. Ratajczak, T. Grajek, "High Efficiency 3D Video Coding using new tools based on view synthesis", IEEE Transactions on Image Processing, Vol. 22, No. 9, September 2013, pp. 3517-3527.

[37] Y. Gao, G. Cheung, T. Maugey, P. Frossard, J. Liang, “Encoder-driven inpainting strategy in multiview video compression”, IEEE Transactions on Image Processing, Volume: 25, 2016, pp. 134 – 149.

[38] P. Merkle, C. Bartnik, K. Müller, D. Marpe, T. Wiegand, „3D video: Depth coding based on inter-component prediction of block partitions”, 29th Picture Coding Symposium, PCS 2012, Kraków, May 2012, pp. 149-152.

[39] K. Müller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F. Hunn Rhee, G. Tech, M. Winken, T. Wiegand, “3D High-Efficiency Video Coding for multi-view video and depth data”, IEEE Transactions on Image Processing, Volume: 22, 2013, pp. 3366 – 3378.

[40] F. Shao, W. Lin, G. Jiang, M. Yu, “Low-complexity depth coding by depth sensitivity aware rate-distortion optimization”, IEEE Transactions on Broadcasting, Volume 62, Issue 1, pp. 94 – 102, 2016.

[41] M. .Hannuksela, D. Rusanovskyy, W. Su, L.Chen, Ri Li, Pa. Aflaki, D. Lan, Michal Joachimiak, H. Li, M. Gabbouj, “Multiview-video-plus-depth coding based on the Advanced Video Coding standard”, IEEE Transactions on Image Processing, Volume: 22, Issue: 9, 2013, pp. 3449 – 3458.

[42] J. Stankowski, Ł. Kowalski, J. Samelak, M. Domański, T. Grajek, K. Wegner, "3D-HEVC extension for circular camera arrangements", 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 3DTV- Con 2015, Lisbon, Portugal, July 2015.

[43] J. Samelak, J. Stankowski, M. Domański, “Adaptation of the 3D-HEVC coding tools to arbitrary locations of cameras”, International Conference on Signals and Electronic Systems, Kraków, 2016.

[44] M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, Ł. Kowalski, M. Kurc, A. Łuczak, D. Mieloch, R. Ratajczak, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, “Methods of high efficiency compression for transmission of spatial representation of motion scenes”, IEEE Int. Conf. Multimedia and Expo Workshops, Torino 2015.

[45] R. De Queiroz, P. Chou, “transform coding for point clouds using a Gaussian process model”, IEEE Transactions on Image Processing, DOI: 10.1109/TIP.2017.2699922, Early Access Article, 2017.

[46] ISO/IEC IS 23009: “Information technology — Dynamic adaptive streaming over HTTP (DASH) “.

[47] T. C. Thang; Q.-D. Ho; J. W. Kang; A. T. Pham, “Adaptive streaming of audiovisual content using MPEG DASH”, IEEE ransactions on Consumer Electronics, vol. 58, 2012, pp. 78-85.

[48] ISO/IEC IS 23008-1: 2013, “Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 1: MPEG media transport (MMT)”.

[49] K. Kim, K. Park, S. Hwang, J. Song, “Draft of White paper on MPEG Media Transport (MMT)”, ISO/IEC JTC1/SC29/WG11 Doc. MPEG-2015 N15069, Geneva, February 2015.

[50] C. Cruz-Neira, D. J.Sandin, T. DeFanti, R. Kenyon, Robert, J. Hart, "The CAVE: Audio visual experience automatic virtual environment". Commun. ACM. 35 (6), 1992, pp. 64–72.

[51] Fraunhofer HHI, “TiME Lab”, www.hhi.fraunhofer.de/en/ departments/vit/technologies-and-solutions/capture/panoramic-uhd-video/time-lab.html, retrieved on April 21, 2017.

[52] A. J. Fairchild, S. P. Campion, A. S. García, R. Wolff, T. Fernando and D. J. Roberts, "A Mixed Reality Telepresence System for Collaborative Space Operation," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 4, pp. 814-827, April 2017.

[53] Holografika,”HoloVizio C80 3D cinema system”, Budapest, http://www.holografika.com/Products/NEW-HoloVizio-C80. html, retrieved on April 21, 2017.

[54] “3D world largest 200-inch autostereoscopic display at Grand Front Osaka”, published: 28 April 2013, https: //wn.com/3d_world_largest_200-inch_autostereoscopic_display at_grand_front_osaka.

[55] NICT News, Special Issue on Stereoscopic Images, no. 419, November 2011.

[56] D. Nam, J.-H. Lee, Y. Cho, Y. Jeong, H. Hwang, D. Park, “Flat Panel Light-Field 3-D Display: Concept, Design, Rendering, and Calibration”, Proceedings of the IEEE, Vol. 105, May 2017, pp. 876-891.

[57] www.oculus.com/rift/ - available April 2017.

[58] https://vr.google.com/cardboard/ - available April 2017.

[59] J. Zaragoza, T. J. Chin, Q. H. Tran, M. S. Brown, D. Suter, "As-Projective-As-Possible Image Stitching with Moving DLT," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1285-1298, July 2014.

[60] M. Z. Bonny, M. S. Uddin, "Feature-based image stitching algorithms," 2016 International Workshop on Computational Intelligence (IWCI), Dhaka, 2016, pp. 198-203.

[61] “Cylindrical Equidistant Projections”, http://mathworld.wolfram.com/CylindricalEquidistantProjection.html

[62] "Cylindrical Projection" http://mathworld.wolfram.com/CylindricalProjection.html

[63] Y. Ye, E. Alshina, J. Boyce, “Algorithm descriptions of projection format conversion and video quality metrics in 360Lib” Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 6th Meeting: Document: JVET-F1003-v1, Hobart, AU, 31 March – 7 April 2017.

[64] “ISO/IEC DIS 23090-2 Omnidirectional Media Format" ISO/IEC JTC1/SC29/WG11 N16824 April 2017, Hobart, Australia

[65] "Requirements for Omnidirectional Media Format" ISO/IEC JTC1/SC29/WG11 N 16773 April 2017, Hobart, Australia.

[66] "Draft Requirements for future versions of Omnidirectional Media Format" ISO/IEC JTC1/SC29/WG11 N 16774 April 2017, Hobart, Australia.

[67] G. Bang, G. S. Lee, N. Ho H., “Test materials for 360 3D video application discussion”, ISO/IEC JTC1/SC29/WG11 MPEG2016/M37810 February 2016, San Diego, USA

[68] K. Wegner, O. Stankiewicz, T. Grajek, M. Domański, “Depth estimation from circular projection of 360 degree 3D video” ISO/IEC JTC1/SC29/WG11 MPEG2017/m40596, April 2017, Hobart, Australia.

[69] M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz, K. Wegner, “Multiview test video sequences for free navigation exploration obtained using pairs of cameras”, ISO/IEC JTC1/SC29/WG11, Doc. MPEG M38247, May 2016.

[70] V. Baroncini, K. Muller, S. Shimizu, “MV-HEVC Verification Test Report” Joint Collaborative Team on 3D Video Coding Extensions of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Document: JCT3V-N1001, 14th Meeting: San Diego, USA, 22–26 Feb. 2016.

[71] V. Baroncini, K. Muller, S. Shimizu, “3D-HEVC Verification Test Report” Joint Collaborative Team on 3D Video Coding Extensions of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Document: JCT3V-M1001, 13th Meeting: Geneva, CH, 17–21 Oct. 2015.