2 Emerging Imaging Technologies: Trends and Challenges

Marek Domański[1] Tomasz Grajek[1], Caroline Conti[2], Carl James Debono[3], Sérgio Faria[4], Peter Kovacs[5], Luís Lucas[4], Paulo Nunes[2], Cristian Perra[6], Nuno Rodrigues[4], Mårten Sjöström[7], Luís Ducla Soares[2], Olgierd Stankiewicz[1]

[1]Chair of Multimedia Telecommunications and Microelectronics, Poznań University of Technology, Poznań, Poland, ({marek.domanski, tomasz.grajek, olgierd.stankiewicz} @put.poznan.pl)

[2]Instituto de Telecomunicações and ISCTE – Instituto Universitário de Lisboa, Lisbon, Portugal (email: {caroline.conti, lds, paulo.nunes}@lx.it.pt)

[3]Department of Communications and Computer Engineering, University of Malta, Msida, Malta, (email: c.debono@ieee.org)

[4]Instituto de Telecomunicações and Politécnico de Leiria, Leiria, Portugal, (email: {sergio.faria, nuno.rodrigues}@co.it.pt, luisfrlucas@gmail.com)

[5]Holografika, Budapest, Hungary, (p.kovacs@holografika.com)

[6]Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy (email: cperra@ieee.org)

[7]Department of Information Systems and Technology, Mid Sweden University, Sundsvall, Sweden, (email: Marten.Sjostrom@miun.se)

Abstract

This chapter addresses image and video technologies related to 3D immersive multimedia delivery systems with special emphasis on the most promising digital formats. Besides recent research results and technical challenges associated with multiview image and image, video and lightfield acquisition and processing, the chapter also presents relevant results from international standardization activities in the scope of ISO, IEC and ITU. Standard solutions to encode multiview image and video content and ongoing research is addressed, along with novel solutions to enable further developments in the emerging technologies dealing with capture and coding for lightfield content and Free Viewpoint Television.

2.1. Introduction

Recently[1], both among the research community and in industry, great attention is paid to *immersive multimedia*. The word *immersive* comes from Latin verb *immergere*, which means to dip, or to plunge into something. In the case of digital media, this term is used to describe the technical systems that are able to absorb viewers totally into an audiovisual scene (Aggoun et al. 2013), (Isgro et al. 2004, Domański et al. 2017). Although i*mmersive multimedia* may be related to both natural and computer-generated content, in this book, we are going to focus mainly on the natural visual content that originates from multiple

[1] Written in 2017.

synchronized video cameras, and that possibly is augmented by data from supplementary sensors, like depth cameras.

For an immersive system, it is important to reconstruct a portion of an *acoustic wave field* (Benesty et al. 2008) and a *lightfield* (Ziegler et al. 2014). In a classic audiovisual system, audio and video are acquired using a single microphone and a single video camera. This 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 and rendered.

Although both video and audio are substantial for the impression of immersiveness, the scope of this book limited to the visual content. Nevertheless, it is worth to mention that significant progress is already made in the immersive and spatial audio technology. The faster development of this audio technology is related to lower bitrates and smaller data volumes for audio than for video. Moreover, the human auditory system is also less demanding than the human visual system. There already 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* (Herre et al. 2015) that are able to produce strong impressions of immersiveness. Firstly, the presentation technology seems to be more advanced for spatial audio than for video. The respective systems comprise the systems with high numbers of loudspeakers but also to the 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 (Blauert 2013) .

During the last decade, the respective spatial audio representation and compression technologies have been developed and standardized in MPEG-D (MPEG Surround 2007), (SAOC 2016) and MPEG-H Part 3 (3D Audio 2015) international standards. The spatial audio compression technology is based on coding 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 (USAC 2016) .

For the *immersive video*, the development is more difficult, nevertheless the research on immersive visual media is booming recently. *Immersive video* (Isgro et al. 2004) may be related to both natural and computer-generated content. Here, we are going to discuss mostly the natural content that originates from video cameras and possibly is augmented with data from supplementary sensors, like depth cameras. Such content is sometimes described as *high-realistic* or *ultra-realistic*. 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. (Ishida and Shibata 2010)).

The above mentioned immersive natural content usually is preprocessed by computers before being presented to viewers. A good example of such *interactive* content is spatial video that allows a viewer to virtually walk through a tropical rainforest reach of hidden swamps, poisonous plants and dangerous animals. During the virtual walk, a virtual explorer is very safe and may enjoy the beauty of nature being relaxed, and without fear. The virtual walker may choose arbitrary a virtual trajectory of a walk, may choose a current direction of view, may stop and look around, watch animals and plants etc.

The respective visual content is acquired with the use of many synchronized cameras. Then, sophisticated computer processing of video is needed in order to produce the entire representation of the visual scene. Presentation of such content usually must be preceded by rendering that results in the production of video that corresponds to a particular location and view direction currently chosen by a virtual rainforest explorer. Therefore, the presentation of such rendered video may also be classified as presentation of *virtual reality* although all the content represents real-world objects in their real locations and motions (see e.g. (EBU 2017)).

Similar effects may be obtained for computer-generated contents, both standalone or mixed with natural content. In the latter case, we 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" (EBU 2017). 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" (EBU 2017).

Considering the immersive video, we have to refer to 360-degree video that is currently under extensive technological development. The 360-degree video allows at least to watch video in all directions around a certain virtual position of a viewer. More advanced versions of 360-degree video allow a viewer also to watch video in any direction up and down from its virtual location, as well as to change the virtual location. In popular understanding, the 360-degree video is even treated as a synonym to the immersive video, e.g. see Wikipedia (WIKI 2017) .

The preliminary classification of immersive video (Domański et al. 2017) was recently discussed by MPEG (Moving Picture Experts Group, i.e. formally ISO/IEC JTC1 SC29 WG11[2]) (OMF 2017), (ISO 2017)). By drawing conclusions from this discussion some main categories of content may be defined:

1. monoscopic 360-degree video, where usually video from many cameras is stitched to a panorama,
2. stereoscopic and binocular 360-degee video that allows a viewer to watch in an arbitrary position with various levels of spatial sensations,
3. 6-degree of freedom 360-degree video that provides a viewer the ability to change freely his/her location.

For Class 2, the first generation of 3D video, i.e. the stereoscopic video is the very popular and the simplest case. The last wave of

[2] See Section 2.3.

enthusiasm for 3D video was encountered around year 2010 but the lack of user-friendly stereoscopic displays has reduced the interests recently. In this book, we rather consider the next-generation 3D content that allow a viewer to perceive spatial parallax possibly without special glasses that are necessary for traditional stereoscopic displays, like shutter glasses, polarization glasses or color-filter glasses. Such glass-free systems are still challenging even for a fixed view, nothing to say about 360-degree video.

The Class 3 is related to virtual navigation that is a functionality of future interactive video services where a user is able to navigate freely around a scene. The systems that provide such functionality are often called free-viewpoint television (FTV) (Tanimoto et al. 2012, Lafruit et al. 2016, Domański et al. 2016, Domański et al. 2015). The prospective FTV will be an interactive internet-based system that may output virtual monoscopic video, virtual stereoscopic video or even multiview video, e.g. for watching a virtual view on an autostereoscopic display.

In 360-degree video, virtual navigation and other types of advanced visual content, the virtual views are synthesized or rendered using a scene representation, or a scene model. The following scene representation types are mostly considered in the references: object-based (Miller et al. 2006, Smolic et al. 2005), ray space (Tanimoto et al. 2012), (Tanimoto 2006), point-based (Wei et al. 2013) , and multiview plus depth (MVD) (Müller et al. 2011) . As the first three types of models are related to quite complex calculations, currently, the MVD representation is used most often and will be extensively considered further in this book. Nevertheless, it is worth to mention that modeling of 3D scenes using point clouds is considered as an competitive and interesting approach, even related to recent standardization projects (ISO 2017a).

The multiview plus depth video format is also vital for the display technology. Although the display technology is also not mature enough for wide adoption of 3D video and for the immersive video and images, the situation seems to be diversified for various display application areas. In particular, the glassless autostereoscopic displays and projection systems are being improved step by step, thus increasing the comfort and quality of spatial (3D) video presentations. Such signage systems may use even 200 views, i.e. they display simultaneously 200 views in order to produce realistic impression of depth (Holografika 2017, Grand Front Osaka 2013, NICT 2011).

2.2. Multiview Video plus Depth

The complete and general description of a visual scene may be provided using a Plenoptic Function (POF) (Adelson et al. 1991). The plenoptic function is usually defined as a function of seven variables, i.e. $POF(x,y,z,\phi,\varphi,t,\lambda)$, where x,y,z represent the coordinates of a point in 3D space, ϕ and φ define the direction of a light ray, t denotes time, and λ denotes the wavelength in light ray. The value of the plenoptic function expresses the "amount of light" (e.g. luminance) of a given wavelength λ, registered at a time instant t at a point (x,y,z), and in the direction defined by the angles ϕ and φ . In order to describe a scene entirely, the plenoptic function should be measured at all points (x,y,z) in some 3D space relevant to the scene, for all wavelengths λ from the

visible light interval, and in all directions defined by the angles ϕ and φ possibly from the interval (-π, π). Obviously, such full description is neither possible nor necessary. Instead, in multimedia technology, we use various simplified representations of 3D scenes already mentioned in the previous Section 3.1. As already mentioned, among those types of representation, the Multiview plus Depth (MVD) representation is the most popular in practical approaches to natural 3D video. More view with the corresponding depth maps we have, more exact is the approximation of the lightfield.

The high number of video views of multiview video results in a huge amount of data that needs to be transmitted over bandwidth-limited channels. This fact motivates the research on compression systems that should be able to drastically reduce the storage and the bandwidth requirements for 3D video data. Practical systems register, process and transmit only a subset of the required views together with the geometric information of the scene, represented by depth maps. The missing views can then be generated at the receiver side through view synthesis algorithms, based on the transmitted v and depth data. For this purpose, depth maps provide the information related to the distance of each pixel in the video view with relation to the view camera position. Such representation for 3D video, using a small number of video views combined with the geometric information of the scene, is the called Multiview Video-plus-Depth (MVD) (Müller et al. 2011, Müller et al. 2013) as already mentioned. Figure 1 illustrates an MVD system, which uses view synthesis at the receiver side.

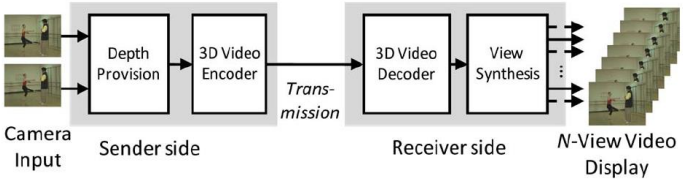


Figure 1 – MVD system based on view and depth data with view synthesis at the decoder side (Mueller et al. 2011)

An example of a depth map and the corresponding view is depicted in Figure 2.

Depth estimation is still a challenging task. In general, there exist two approaches:

application of special depth sensors called also depth cameras (e.g. Stamos and Allen 2000, Sandberg et al. 2011),

- estimation of depth from video data by the use of video analysis on computers.

The depth sensors illuminate a scene with invisible infrared light and mostly exploit one of the following two technologies:

- by measurements of the time-of-flight (Gokturk et al. 2004) from the radiator to the object and back to the sensor,
- by analysis of structured light reflected from a scene illuminated with a specific pattern.

Currently, both technologies are under further development resulting in their improvements. Despite of which technology is used, the usage of depth sensors is conceptually very attractive as they produce may produce the depth in real time with reasonable latency. Nevertheless, their practical employment still faces severe problems related to limited spatial and temporal resolutions of the acquired depth maps, limited distance ranges, synchronization of video and depth cameras, additional infrared illumination of the scene that may interfere with other equipment, mutual interference of several sensors working simultaneously at the same scene, and sensitivity to environmental factors including solar illumination. Currently, these sensors are only capable of acquiring low-resolution depth maps, which are usually enhanced by post processing methods based on interpolation and denoising filters. Also, the maximum and minimum depth value acquired by these sensors is limited. Furthermore, since depth sensors are physically independent of video cameras, they are positioned at slightly different positions, resulting in depth maps that do not exactly match the associated views. Already, substantial research work is done with the aim to overcome the abovementioned problems, see e.g. (Kang and Ho 2010, Sen et al. 2013, Wang 2015). Despite of all the abovementioned problems, the technology of depth cameras is intensively developed for many potential applications including industrial computer vision, mobile robot navigation, control of autonomous cars, and many others.

- A view and the corresponding depth map for the camera 5 of *Shark* test multiview sequence.

Depth estimation is still a challenging task. In general, there exist two approaches:

- application of special depth sensors called also depth cameras (e.g. Stamos and Allen 2000, Sandberg et al. 2011),
- estimation of depth from video data by the use of video analysis on computers.

The depth sensors illuminate a scene with invisible infrared light and mostly exploit one of the following two technologies:

- by measurements of the time-of-flight (Gokturk et al. 2004) from the radiator to the object and back to the sensor,
- by analysis of structured light reflected from a scene illuminated with a specific pattern.

Currently, both technologies are under further development resulting in their improvements. Despite of which technology is used, the usage of depth sensors is conceptually very attractive as they produce may produce the depth in real time with reasonable latency. Nevertheless, their practical employment still faces severe problems related to limited spatial and temporal resolutions of the acquired depth maps, limited distance ranges, synchronization of video and depth cameras, additional infrared illumination of the scene that may

interfere with other equipment, mutual interference of several sensors working simultaneously at the same scene, and sensitivity to environmental factors including solar illumination. Currently, these sensors are only capable of acquiring low-resolution depth maps, which are usually enhanced by post processing methods based on interpolation and denoising filters. Also, the maximum and minimum depth value acquired by these sensors is limited. Furthermore, since depth sensors are physically independent of video cameras, they are positioned at slightly different positions, resulting in depth maps that do not exactly match the associated views. Already, substantial research work is done with the aim to overcome the abovementioned problems, see e.g. (Kang and Ho 2010, Sen et al. 2013, Wang 2015). Despite of all the abovementioned problems, the technology of depth cameras is intensively developed for many potential applications including industrial computer vision, mobile robot navigation, control of autonomous cars, and many others.

Depth can be also estimated in the process of video analysis. The real views used for depth estimation should be corrected by compensation of the lens distortions, and possibly also by compensation of the differences in color characteristics of the cameras. Moreover, illumination differences also should be compensated.

The depth estimation may be described as follows. For the simplest case, consider two views. The pairs of characteristic points need to be found in the views. For each such pair, disparity d can be measured as the shift between the locations of the corresponding characteristic in the two views. Assume that the focal length of both cameras is f, and the distance between the optical centers of the cameras, i.e. the base distance is b. Assuming $f \ll z$ we get (Hartley and Zisserman 2015), we may calculate the depth of a point object

$$z = \frac{fb}{d} \ . \qquad (1)$$

In order to use Formula 1, the values of focal length f and the base b need to be measured. It is done in the process of calibration of the multi-camera system, when some special calibration video is recorded, and the relevant camera parameters as well locations of the camera sensors are estimated using the data obtained from the calibration video (Zhang et al. 2000).

Estimation of depth from a pair of views has been studied since many years (e.g. Atzpadin et al. 2004, Lee and Ho 2010, Min et al. 2010) . Some methods (Bleyer and Gelautz 2005, Hang and Chen 2004) focus on the segmentation-aided depth estimation based on optimization performed on a graph. While achieving relatively high quality of estimated depth maps, these methods are designed for stereo pairs only. Moreover, main optimization process is performed on the pixel level, making the whole estimation very time-consuming. Exploitation of the outputs from more than 2 cameras provides the opportunity to produce more exact depth maps. For example, the method of (Zilly at al. 2014) estimates depth maps for limited resolution in the real-time, using the outputs from 4 cameras with parallel optical axes. The method of (Jorissen et al. 2015)] proposes the estimation of the multiview depth based on the epipolar plane image. While providing the inter-view consistent depth of the high

quality, this method is still limited to linear arrangements of cameras. Multiview depth estimation can be based on the Belief Propagation (Sun et al. 2003). In the work describe in (Montserrat et al. 2009], the inter-view consistency is ensured by depth maps cross-checking and multiview matching of views. The methods have been also proposed that provide the temporal consistency of the estimated depth maps (Stankiewicz et al. 2015), (Mieloch et al. 2017). There exist a huge number of papers on various aspects of the depth estimation, and this paragraph provides sparse samples of the references rather than an entire review.

The Depth Estimation Reference Software (Stankiewicz et al. 2013) has been developed by MPEG, and currently, it is widely used a reference for multiview depth estimation.

Recently, it was shown that for highly-occluded scenes nonuniform distribution of cameras around a scene leads to better depth estimation (Domański et al. 2016a). Therefore, for such real scenes, it was proposed to acquire multiview video using camera pairs (Domański et al. 2016).

Obviously, depth maps can be represented as greyscale images. In practice, the name of depth map is used for the data sets, where the samples represent either depth or disparity. The depth or disparity samples have often 8-bit representation. If disparity representation is used, each sample value corresponds to inverse of the distance from the given camera to a given scene point, or more exactly to the plane that contains this particular scene point and is perpendicular to the optical axis of the camera. It means that the range between the minimum and maximum depth distances is divided into 256 unequal intervals. Closer distances are represented more accurately while the further ones more sparsely. Therefore, for many applications, the depth sample representations with more than 8 bits are used.

Depth estimation allows to produce the multiview plus depth representation that may be used for synthesis of virtual views, or, in other words, for Depth-Image-Based Rendering (DIBR) that is essential for free viewpoint television, augmented and virtual reality, lightfield displays etc. The virtual view synthesis is also exploited in order to increase compression efficiency for multiview video (Domański et al. 2013).

Figure 2 presents a block diagram of the DIBR algorithm, based on two reference views and their associated depth maps. Any virtual view can be generated based on these two references. Usually two nearest real views, labelled left and right reference views in Figure 2, are selected from the multi-view sequence and warped (Tanimoto 2011). The warped images generated from the two views are then blended to form the new virtual position (Do et al. 2009). Since some disoccluded regions and holes may still remain, inpainting is applied to fill the missing data (Tanimoto 2011).

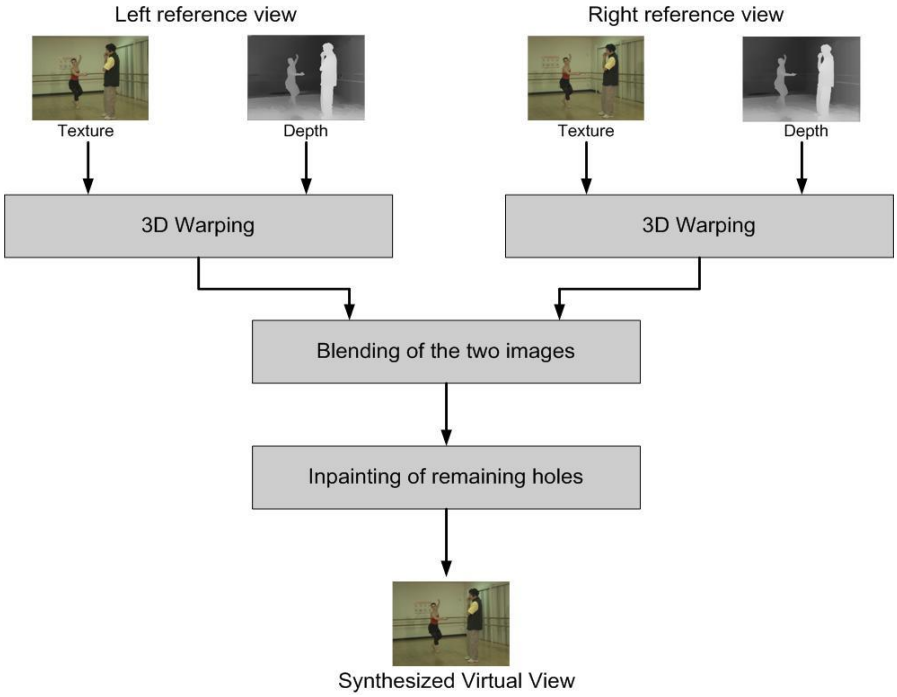


Figure 2 - Block diagram of the DIBR algorithm.

In order to reduce errors introduced by stereo matching algorithms, (Oh et al. 2010) proposes a depth map pre-processing algorithm based on temporal filtering, compensation for errors and spatial filtering. An illumination compensation technique is applied in (Yang et al. 2011) to reduce colour discontinuities and improve visual quality of the synthesized views. The warped depth maps are processed by median and bilateral filters before inverse warping in (Wegner et al. 2013) to improve the visual quality of the synthesized view. Furthermore, depth map pixels at edges are detected and are not warped in (Zarb and Debono 2014). This technique reduces unreliable data in these regions from the warping operations.

Other DIBR techniques found in literature include the enhancement of virtual views through pixel classification, graph cuts and depth-based inpainting (Tran and Harada 2013) . The perceived depth quality and visual comfort in stereoscopic images are improved using stereoacuity before rendering the images in (Xu et al. 2014). Furthermore, a just noticeable depth difference (JNDD) model and saliency analysis is used in (Lei et al. 2015) to provide a better user perception of the rendered content. Recently, good-quality synthesis technique has been demonstrated for practical virtual navigation in a scene represented by multiview plus depth with real cameras sparsely located around a scene (Dziembowski et al. 2016). For research purposes, the View Synthesis Reference Software (Wegner et al. 2013) is available in the version adequate for synthesis of the views from arbitrary locations.

The data processing pipeline for multiview plus depth representation of visual scene together with the corresponding audio data is depicted in Figure 4.

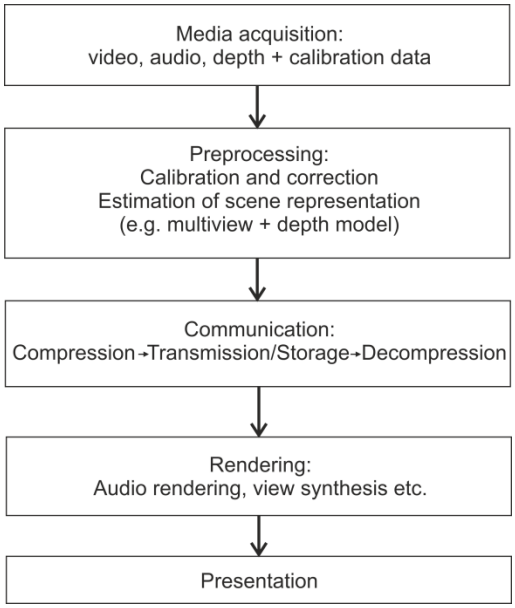


Figure 4. The processing chain for spatial video associated by spatial audio (Domański et al. 2017)

2.3. Standardization – the status and current activities

2.3.1. Standardization in multimedia

Standardization is crucial for telecommunications where the transmitter and the reciever are often placed in the locations being very distant one from the other. In such cases, the interoperability of hardware and software delivered by different vendors is an issue of paramount importance. The means to ensure the interoperability is to observe standards agreed by all involved paties. In practice, such standardization agreements are obtained either in international instititutions or by consortia of companies sharing substantial portions of the relevant markets.

The following international institutions play the primary role in multimedia standardization:

ISO – International Organization for Standardization,

IEC – International Electrotechnical Commission,

ITU – International Telecommunication Union.

In the area of multimedia, ISO and IEC work mostly jointly and they jointly issue international standards (IS). International standards are therefore numbered as, e.g. ISO/IEC IS 14496. Except of the number, each standard has also its own generic name. The ISO/IEC standards are devided into parts, like Part 1 “Systems”, Part 2: Video, Part 3 “Audio” etc. In fact, a part of the standard defines the minimum requirements for interoperability for a given tehnology, like video compression or audio compression. The parts of the standards may also be recommendations of ITU. The standards (called recommendations) of ITU are grouped into Telecommunication Sector (ITU-T) and Radiocommunication Sector (ITU-R). Of course, some standards are independently developed and issued by only one

institution, some are issued jointly by two or hree of them. Moreover, some internationally recognized standards have been also defined by IEEE, i.e. The Institute of Electrical and Electronics Engineers and by SMPTE (Society of Motion Picture and Television Engineers).

Moreover, there also regional and national standardization organization. For example, the Chinese consortium for Audio Video Coding Standard plays importnant role in standardization of the compression of video and audio.

In many cases, the active role is played by an industrial consortium. For example, a group of big companies (Amazon, ARM, Cisco, Google, Intel, Microsoft, Mozilla, Netflix, NVidia) has recently created an Alliance for Open Media with the aim of producing a new standard for video compression called AV1.

For video and audio compression, the minimum interoperability requirements are related to the semantics and syntax of the bitstream, i.e. they define how to read the bitstream. It means that a standard defines the decoders, while having limited impact on theencoders (cf. Figure 5).

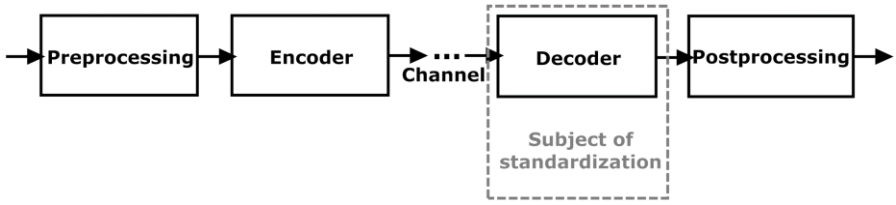


Figure 5. Standardization of compression

2.3.2. Basic technologies

In the recent years, significant effords have been made in standardization of compression of multiview video, multivie plus depth video as well as other related aspects. These techniques mostly rely on the consecutive generations of monoscopic video coding. During the last 25 years, consecutive generations of monoscopic video coding technology have been accepted as the international standards, like MPEG-2 (MPEG-2 2012), AVC (Advanced Video Coding) (AVC 2014), and HEVC (High Efficiency Video Coding) (HEVC 2016). Currently, the new generation of video compression technology is under development and is expected to be standardized around 2020-2021 as a part of the prospective MPEG-I standard. These consecutive video coding generations have been developed thanks to huge research efforts.

For a given content type, and for a given video format, the bitrate of the compressed bitstream may be very roughly assessed assuming the required quality level and a mature codec implementation that reaches nearly the same compression as the standard reference software. For demanding complex dynamic content and assumed broadcast quality, for monoscopic video codecs the bitrate *B* may be very roughly estimated using the formula (Domański et al. 2012, Domański 2015, Domański et al. 2015a)

$$B \approx A \cdot V \quad [\text{Mbps}] , \tag{2}$$

where *A* is technology factor, where

A = 4 for MPEG-2,
A = 2 for AVC,
A = 1 for HEVC,
A = 0.5 for the prospective technology expected around year 2021,
and V is video-format factor, where
V=1 for the Standard Definition (SD) format, (either 720×576, 25 fps or 720×480, 30 fps, chroma subsampling 4:2:0, i.e. one chroma sample from each chroma component C_R and C_B per 4 luma samples),
V=4 for the High Definition (HD) format (1920×1080, 25/30 fps, chroma subsampling 4:2:0),
V=16 for the Ultra High Definition (UHD) format (3840×2160, 50/60 fps, chroma subsampling 4:2:0).

The conceptually simplest way to implement the coding of multiview video is to encode each view as an independent video stream with the time stamps included. Such type of compression is usually called simulcast coding. Simulcast coding exploits the commonly used relatively cheap video codecs may be efficiently applied. The total bitrate B_m of the bitstreams is

$$B_m = N \cdot B, \tag{3}$$

where N – the number of views,
B – the bitrate for a single view from Equation 2.

2.3.3. Multiview video coding

The main idea of the multiview video coding is to exploit the similarities between neighboring views. One view, called the base view, is encoded like a monoscopic video using standard intraframe and temporal interframe predictions. The respective bitstream constitutes the base layer of the multiview video representation. The base view may be decoded from the base-layer bitstream using a standard monoscopic decoder. For encoding of the dependent views, i.e. the other views the inter-view prediction with disparity compensation may be used in addition to standard intraframe and interframe predictions. In inter-view prediction, a block in a dependent view is predicted using a 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, where the motion vectors are replaced by the disparity vectors.

In multiview video coding, the pictures are predicted not only from temporal interframe references, but also from inter-view references. An example of a prediction structure is shown in Fig. 6.

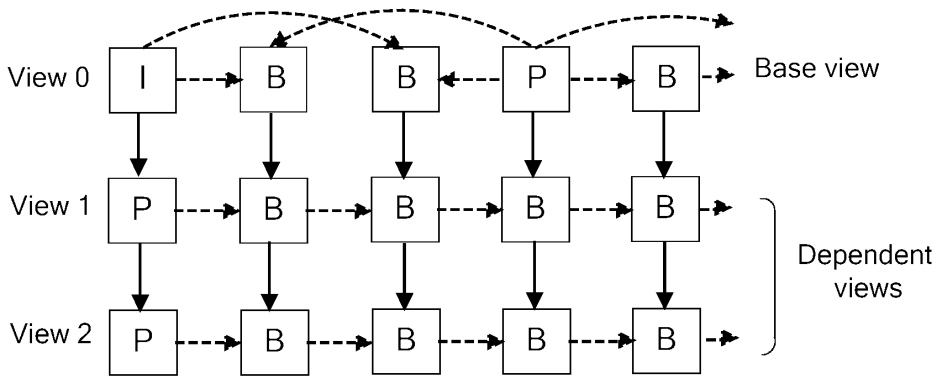


Figure 6. Typical frame structure in multiview video coding using inter-view prediction with disparity compensation: solid line arrows denote inter-frame predictions while dashed line arrows correspond to temporal predictions. The letters I, P, and B denote I-frames (intraframe coded), P-frames (compressed using intra- and temporal interframe coding) and B-frames (compressed using two reference frames).

Multiview video coding has been already standardized as extensions to the MPEG-2 standard (Haskell et al. 1996), the AVC standard (Vetro et al. 2011), and the HEVC standard (Tech et al. 2016). The multiview extension of AVC is denoted as MVC, and that of HEVC as MV-HEVC. These multiview extensions have been standardized in such a way that only minor modifications are needed to the monoscopic codec implementations. Therefore some more advanced techniques for multiview coding are not included into the standards.

For the state-of-the-art multiview video coding technology is MV-HEVC (HEVC 2016).

The multiview coding provides the bitrate reduction of order 15-30%, sometimes reaching even 50% as compared to the simulcast coding. 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.

Recently (Samelak et al. 2017), it was shown that the efficiency of the inter-view prediction is virtually the same for Multiview HEVC and for HEVC augmented by Intra Block Copy tool (originally designed for computer-generated content) using the same resolution of translation/displacement vectors. It is worth to add that the latter codec has simpler single-loop structure and is nearly compliant with standard HEVC Screen Content Codec. The result was obtained for rectified multiview video clips acquired using cameras with parallel optical axes, i.e. for the application scenario, for which Multiview HEVC was designed. This result put in question the need to develop multiview video codecs for future generations of video compression techniques.

2.3.4. 3D video coding

Many 3D video coding tools have been already proposed including prediction based on: view synthesis, inter-view prediction by 3D mapping defined by depth, coding of disoccluded regions, advanced inpainting, special techniques for depth coding using platelets and wedgelets etc. (Chen et al. 2016), (Domański et al. 2013), (Gao et al. 2016), (Merkle et al. 2012), (Müller et al. 2013), (Shao et al. 2016). Some of these tools have been already included into the standards of 3D video coding: 3D High Profile of AVC (AVC), (Hannuksela et al. 2013) and 3D Main Profile of HEVC (HEVC), (Tech et al. 2016). The latter defines the state-of-the-art technology for compression of 3D video with accompanying depth. This technology not only compress the depth but also exploits the depth in order to improve coding performance of the multiview video.

For 3D-HEVC, the standardization requirement was to reuse the monoscopic decoding cores for implementations. MV-HEVC, 3D-HEVC, and the scalable extension of HEVC share nearly the same high-level syntax of the bitstreams, and the multi-loop structure of the encoders and decoders is the common architecture used in the implementations. Therefore, view encoding cannot depend on the corresponding depth. As compared to MV-HEVC, 3D-HEVC provides additional prediction types:

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 MV-HEVC is expressed by 2-12% bitrate reduction. Nevertheless, the compression gains of both 3D-HEVC and MV-HEVC 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 simulcast falls below 15%, often being around 5%. This observation stimulated research on the extensions of 3D-HEVC that use true 3D mapping for more efficient inter-view prediction (Stankowski et al. 2015), (Samelak et al. 2016). 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 systems (Domański et al. 2015a).

3D video coding is currently a research topic for several groups around the world, and also future standardization activities are expected. Recently, the MPEG-FTV, the body within MPEG, was exploring possible 3D-HEVC extensions for efficient coding of multiview video taken from arbitrary camera positions. Currently this activity has been shifted to MPEG-I project. Hitherto practical deployment of 3D-HEVC is negligible but growing interests in the applications hitherto mentioned in this chapter will stimulate applications of this compression as well as, probably, standardization of its more efficient extensions. It is also expected that the coding tools of 3D-HEVC together with possible improvements will be included, probably with some delay, into the forthcoming video coding standard that is expected to be ready around 2020-2021 in its first version.

In general, depth maps are characterized by homogeneous regions separated by sharp edges at object boundaries. Despite the distinct

characteristics, multi-view video and depth maps represent the same scene. Therefore video and depth map of a given view exhibit some correlation. The similarities between both streams can thus be exploited by the video coding methods. In such a scheme a base view still needs to be encoded independently from other views and depth maps, allowing compatibility with legacy single-view displays. All the remaining views and depth maps will depend on this stream.

In the scope of the MVD coding, the ISO/IEC MPEG standardization process comes out with three solutions based on different coding technologies. MVC+D is proposed as a simple solution for sending views along with corresponding depth maps, using the Multiview Video Coding (MVC) (Chen et al. 2008) algorithm. All changes are related to high level syntax elements only providing a way to signal the presence of depth data (Chen et al. 2013) . Other two solutions incorporate specific tools for the independent compression of depth maps or for the joint compression of video and depth, based on Advanced Video Coding (AVC) (AVC 2014) and High Efficiency Video Coding (HEVC) (HEVC 2016) encoders. The first one is 3D-AVC algorithm, which is backward compatible with AVC and provides a fast and easy adoption of 3D video in the market. The other one is the current state-of-the-art solution for 3D video coding, known as 3D-HEVC (Müller et al. 2013, HEVC 2016).

The different features of depth maps, associated to the fact that they are not displayed at the decoder, imply that compression of depth maps with the standard video encoder might not be optimal. In order to improve the coding efficiency of depth maps, and the quality of the synthesized views, it has been shown that preservation of depth map edges is very important. In this context, alternative methods based on different coding paradigms have been proposed outside of the scope of standardization groups. The Platelet and Wedgelet depth modelling, pattern-matching coding and linear-fitting modelling are some of the solutions suggested in literature (Wegner et al. 2014, Merkle et al. 2015, Graziosi et al. 2010, Lucas et al. 2015) .

Techniques to save even more bandwidth include the down-sampling of the depth maps. These will require the up-sampling of the decoded maps at the receiver side. In any case, preservation of the edges in the depth maps is very important for view synthesis. Thus, a joint video/depth edge-based up-sampling method can be applied as in (Deng et al. 2012) to better define the edges in the depth map. This is possible because the edges in the depth map are also present in the video, which corresponds to the same scene and objects.

Compression efficiency can be improved by removing high frequency components from both views depth maps. Each image may be divided into regions based on their depth values. Regions which are far away from the camera are low-pass filtered more coarsely than closer regions. This ensures that the removal of the detail does not severely degrade the quality of the image (Aflaki et al. 2014) . This method assumes that the viewer is more concerned with the foreground than with the background. Similarly regions of image further from the camera can be quantized more than closer regions (Domański et al. 2012) . Objects in the view and the depth map video streams move with similar direction and speed. This correlation can be exploited using a Scalable Video Coding (SVC) architecture, where the base layer encodes the views and the enhancement layer carries the

depth data. This idea is presented in (Zhang et al. 2010) and (Tao et al. 2009) where an inter-view prediction scheme is coupled with an inter-layer motion prediction method. The inter-layer motion prediction is based on SVC. Currently this approach is part of a 3D-HEVC standard where depth maps motion field can be predicted from corresponded motion field.

Although MVD requires the additional compression of depth information, it saves a high amount of bits by transmitting a reduced set of views. Furthermore, due to its characteristics, depth maps tend to result in a much smaller compressed bitstream when compared to the video. At the decoder side, a higher number of views can be generated using a synthesis algorithm. One of the most popular techniques is depth-image-based rendering (DIBR) in which the depth data and the view are used to generate the virtual image. This technique was selected by the motion picture experts group (MPEG) as the reference synthesis framework for free-viewpoint video architectures, which relies on the multi-view video-plus-depth format. In fact, the view synthesis reference software (VSRS) that was released by the ad hoc group on 3D audio and visual (3DAV) of MPEG is based on DIBR (ISO 2010) . Although originally it was design only for linear view arrangement, recently it was generalized to cope view general view arrangements as well (Wegner et al. 2013).

2.3.5. New standardization projects

The international organizations work by their working groups of experts. For ISO and IEC there are two groups: JPEG (official name is ISO/IEC JTC1/SC29/WG1) and MPEG (official name is ISO/IEC JTC1/SC29/WG11). For ITU the relevant working group is VCEG. In order create a new general video coding standard that will correspond to more modern compression technology, both ISO/IEC and ITU have created a joint group called Joint Video Exploration Team (JVET). This group is working towards a new video coding standard that will be related to the technology that hales the bitrates of HEVC. Within ISO?IEC this standard will be a part of the forthcoming MPEG-I (from immersive) standard.

In 2017, the MPEG-I standardization project comprises also the extensive works on point cloud compression and lightfield video compression. The latter is also a work item for JPEG that has created already a working subgroup JPEG PLENO that is dealing with lighfield image compression. The lightfield images will be considered in the next two sections.

2.4. Lightfield Super-Multiview with Camera Array

In order to visualize a lightfield, it first needs to be captured. Regardless of the parametrization we, the light field should ideally be captured on a sufficiently large plane, and with the smallest possible granularity both in the spatial and angular sense.

While it is possible to capture light-field using a single sensor (as described in the next section), the physical baseline (distance between the leftmost and rightmost captured position) is limited by the

physical size of the camera. This means that the viewing angle (Field of View) of the captured imagery can only be relatively small, unless the camera is capturing an object from close up. See Fig X. If we are about to capture larger scenes with a large field of view, we either have to use very big cameras (which do not exist in practice), or a camera array spanning the necessary baseline.

It is important to note that while the ultimate goal is to capture a (near) continuous light field, camera arrays can only capture a light field with a specific granularity due to the gap between adjacent cameras. That is, all these camera arrays are sampling the light field at regular intervals, which needs to be taken into account when working with the captured data.

The layout of camera arrays can be quite different depending on the scene and capture requirements. Some special cases include linear, converging linear, and arc setups. Camera arrays can also be 1D or 2D arrays. In a Linear array, cameras are positioned next to each other, with equal distance between cameras, their optical axis is parallel, and perpendicular to the line on which cameras are arranged. In a converging linear array, the position of the cameras is similar, but they are rotated, so that their optical axis points towards a common point of convergence. In case of an arc / circular camera array, cameras are positioned on a circular path, all pointing to a point of convergence in the center. In case of a 1D array, the cameras are arranged in a single row (or column, but that's a quite unusual setting), while in case of a 2D array, cameras are arranged both horizontally and vertically. The advantage of regular camera arrays is that the rough position and orientation of cameras is known, which is later refined by a camera calibration process. Apart from this, an unstructured array of cameras that capture the same scene from different angles can be considered a camera array, and can be used for light-field capture, however the density of the captured data may vary over the field of view.

There are many examples of camera arrays in both research and industrial settings, used for a variety of purposes. A quite well known and early camera array is the Stanford Multi-Camera Array (Wilburn et al. 2005), consisting of 128 video cameras. These cameras can be arranged in various layouts, such as a linear array of parallel cameras having horizontal and vertical parallax, or a converging array of cameras. This large rig has been used for capturing light fields for research purposes, for example for light-field rendering, synthetic aperture imaging. Numerous other multi-camera rigs are known, such as the 100 camera array at Nagoya University (Tanimoto et al. 2005), the 27-camera array at Holografika (Balogh and Kovacs 2010) or the recent horizontal and vertical parallax 16-camera system from Fraunhofer IIS (Zilly et al. 2015). These camera systems provide sufficient input to 3D light-field displays, as the density (in terms of angular resolution) and width (in terms of baseline) of the captured light field allows for wide-angle visualization.

The main design constraint of camera arrays is the physical size of cameras and lenses, which pose an upper limit on how dense the arrangement of such cameras can be. For this reason, typically small camera modules are preferred, while some designers even use board level cameras to achieve an even narrower size per camera.

Using cameras with the possibility of triggering ensures that the frames captured by the individual units represent the same time

instant, which is important to ensure consistency between images when capturing a moving scene.

Static scenes can of course be captured without strict synchronization. Going further, as a special case of a camera “array” one can use a moving camera to capture a static scene (Kim et al. 2013) , or a static camera with a rotating object (Jones et al. 2007) to obtain a light field. These approaches work properly as long as the static scene indeed remains static during the capture session (for example, no people walking by, no changes in illumination due to different position of the sun), and that camera positioning is precise enough to assume that no further camera adjustments are necessary.

Calibration of the individual cameras (resulting in intrinsic camera parameters) is just as important as with single camera capture, however in the case of many-camera arrays, the relative position of cameras (resulting in extrinsic camera parameters) is just as important. Cameras pairs are typically calibrated by using stereo calibration techniques (Zhang 2000) (which can be performed for multiple camera pairs if they can see the same calibration patterns / features), and finding globally consistent extrinsic multi-camera parameters by using an optimization algorithm on the camera parameters (Bo et al. 2013) .

Any kind of regular 2D cameras can be used to build a camera array. For video capture, typically machine vision cameras with trigger capabilities, or professional video cameras are used. DSLRs have also been used for capturing both static and animated light fields. Camera arrays built of GoPro cameras have also been used, however these cameras do not allow for real-time streaming of the captured video over a cable connection – in such cases the recoded light-field needs to be downloaded after the capture session. In real-time light-field capture settings however, the bandwidth required for transfer and store the resulting video data can be a concern.

Researchers not in the possession of a camera array wishing to do research on light-fields can do so using the many available public datasets. A good collection of these can be found in the MPEG FTV Call for Evidence (ISO 2015), which lists selected Super-Multiview and Free Viewpoint Television content to be used for experimentation.

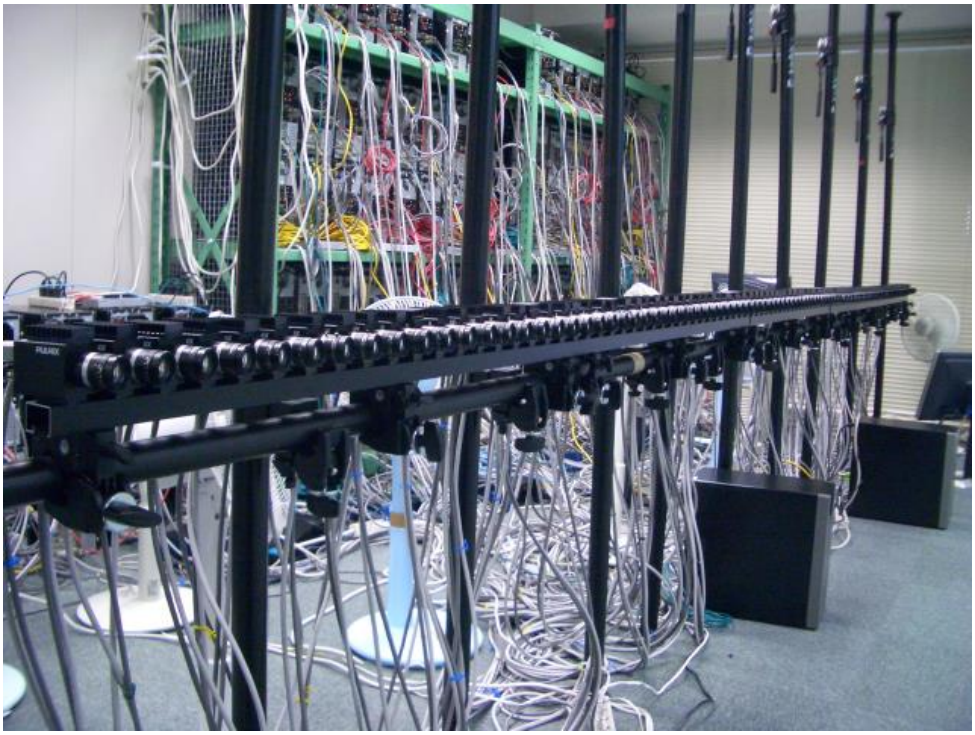


Figure 7: 100 camera array of Nagoya University. (source: Masayuki Tanimoto)

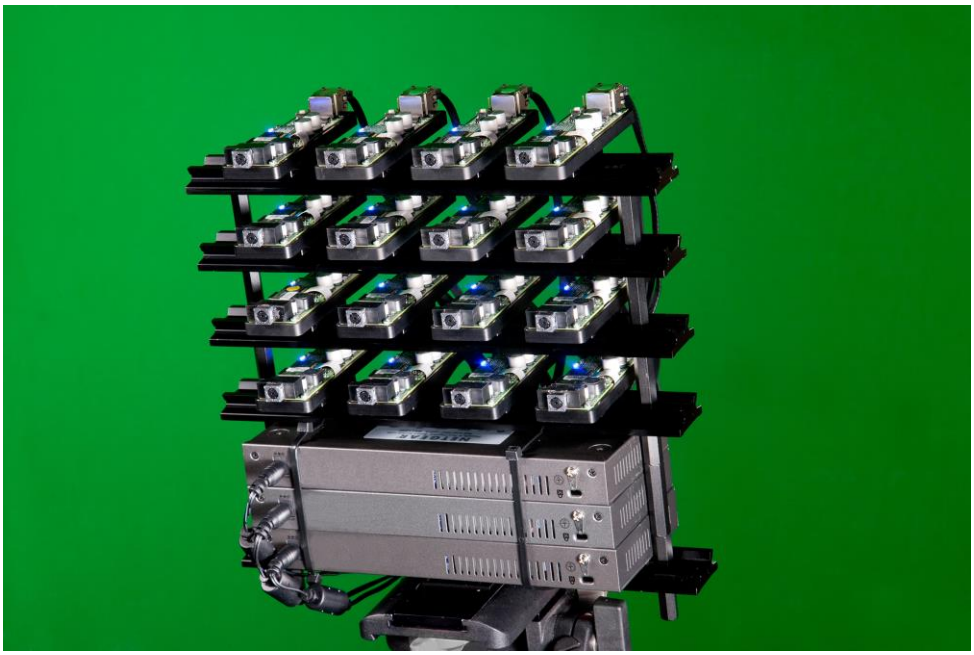


Figure 8: 16-camera full parallax camera array of Fraunhofer IIS. Copyright: Kurt Fuchs| Fraunhofer Institute for Integrated Circuits IIS

Figure 9: 36-camera matrix at Poznan University of Technology, Multimedia laboratory.

2.5. Lightfield with Microlens Array

Lightfield with microlens array – also known as holoscopic (Aggoun et al. 2013) plenoptic (Georgiev and Lumsdaine 2010) and integral imaging (Xiao et al. 2013) – derives from the fundamentals of lightfield/radiance sampling (Levoy and Hanrahan 1996), where not only the spatial information about the 3D scene is represented but also angular viewing direction, i.e., the "whole observable" scene.

The concepts behind this lightfield imaging technology were firstly proposed by G. M. Lippmann and referred to as integral photography in 1908 (Lippmann 1908). The conventional lightfield imaging system comprises a main lens and a regularly spaced array of microlenses, known as a "fly's eye" lens array (Aggoun et al. 2013) which is overlaid with the image sensor at the focal distance, f, as seen in Fig. 10. Therefore, different from a conventional camera that captures an image by integrating the intensities of all rays (from all directions) impinging each sensor element, in a lightfield camera each sensor element collects the light of a single ray (or of a thin bundle of rays as depicted in Fig. 10) that converges on the microlens from a given angular direction.

The traditional lightfield camera can be generalized to alternative camera setups, such as the one proposed in (Georgiev and Lumsdaine 2010) and referred to as focused setup camera. In the focused camera, the main lens and the microlenses are focused in an image plane in front (or behind) of the microlens array plane. As a result, the main lens forms a relay system with each microlens. In practice, these differences in the optical geometry will only change the trade-off

between providing maximal angular or spatial resolution in the captured lightfield image (Ng 2005).

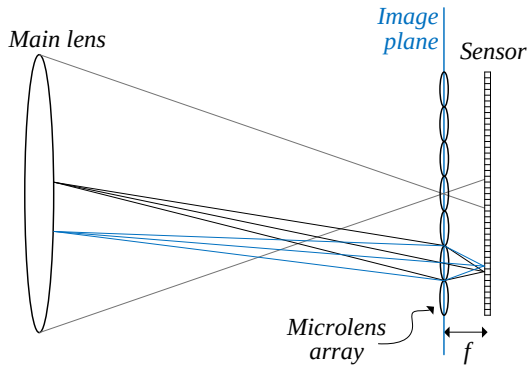


Fig. 10: Basic optical setup of the traditional lightfield camera comprising a main lens, a microlens array and an image sensor

Among the advantages of employing a lightfield imaging system with microlens array is the ability to open new degrees of freedom in terms of content production and manipulation, supporting functionalities not straightforwardly available in conventional imaging systems, namely: post-production refocusing, changing depth-of-field, and changing viewing perspective. Moreover, the interaction functionalities can also be enriched, for instance, by allowing the user to vary the plane of focus and depth of field interactively. In addition to this, it is still possible to derive from this type of content geometric information, such as depth/disparity and ray-space (Tanimoto et al. 2012) representations.

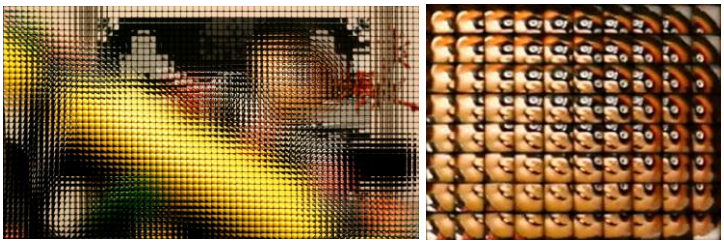
Recently, lightfield imaging with microlens array has become a promising approach for 3D imaging and sensing, being applied in many different areas of research, e.g., 3D television, (Aggoun et al. 2013, Arai 2014) image recognition and medical imaging (Xiao et al. 2013). For this reason, novel initiatives on image and video coding standardization have also considered lightfield application scenarios. Notably, the JPEG working group started recently a new study activity – known as JPEG Pleno (Ebrahimi 2015) – targeting richer image capturing, visualization, and manipulation. In addition, the MPEG group started the third phase of Free-viewpoint Television (FTV), in August 2013, targeting SMV, free navigation and full parallax imaging applications (Tehrani et al. 2013).

However, introducing lightfield image and video applications with its appealing functionalities will require to identify the requirements and challenges in this type of systems, as well as to understand the users' needs in terms of lightfield content interaction. Regarding the challenges, to provide a lightfield representation with convenient spatial resolution and viewing angles, a huge amount of data is required and thus efficient coding is of utmost importance. In addition to this, as the imaging technology moves toward richer representations, novel data representations are essential to support the new applications and functionalities that arise (Ebrahimi 2015). In this sense, a scalable coding architecture is desirable to support a very flexible scaling of the lightfield content with a diverse range of consumption environments and devices. Moreover, this makes it

possible to accommodate in a single compressed bitstream a variety of sub-bitstreams appropriate for users with different preferences and various application scenarios: from the user who wants to have a simple 2D version of the lightfield content without actively interacting with it; to the user who wants full immersive and interactive 3D lightfield visualization. Additionally, providing supplementary data – such as disparity/depth, ray-space, and 3D model – to be incorporated into the scalable bitstream is also important to support lightfield applications that are adaptable to various display interfaces, e.g., stereo, multiview, super-multiview, and also lightfield displays. Hence, it would facilitate the support for displays with different sizes, and with larger number of viewpoints and angular resolutions. Another important requirement is to provide backward compatibility with the current state-of-the-art in image and video coding technologies so as to support interoperability with the widely used 2D and 3D representation formats (Ebrahimi 2015).

Towards the goal of identifying more powerful lightfield representation and coding solutions, several image and video coding schemes have been recently proposed in the literature for the lightfield with microlens array case and try to take advantage of its characteristic planar intensity distribution to achieve more efficient compression. Notably, as a result of the used optical system, the lightfield raw image corresponds to a 2D array of micro-images (MIs), also known as elemental images, where both light intensity and direction information are recorded, as illustrated in Fig. 11a. Due to the small baseline between adjacent microlenses used in the lightfield acquisition process, a significant cross-correlation exists between neighboring MIs (see Fig. b).

In terms of the possible different ways to organize the lightfield data for coding and transmission, the following three main approaches can be identified.



(a) (b)

Fig. 11 Lightfield image captured with a focused setup camera using a 250 μm pitch microlens array: (a) Full image with resolution of 1920×1088; (b) Enlargement of 280×224 pixels showing the array of micro-images

2.5.1. Lightfield raw data-based approach

This category corresponds to cases in which encoding and transmission of the lightfield image is done in its entirety, represented as a 2D grid of MIs. For this, a special lightfield prediction scheme is introduced in a state-of-the-art 2D codec to exploit the non-local spatial redundancy between different MIs for improving the coding efficiency. Fig. 1212 illustrates a basic coding diagram based on the

High Efficiency Video Coding standard (HEVC) (Sullivan et al. 2012) for introducing a lightfield prediction scheme.

Following this approach, a scheme for displacement intra prediction, referred to as self-similarity (SS) estimation and compensation, was proposed in (Conti et al. 2011) to improve the performance of the H.264/AVC standard for lightfield image coding. Later, in (Conti et al. 2012, Conti et al. 2016), the authors proposed to introduce the SS prediction into the HEVC standard for image and video coding so as to take advantage of the flexible partition patterns used in this type of video codecs. In (Bishop and Favaro 2009), the authors investigate alternative non-local spatial prediction, and also propose to include a prediction framework based on locally linear embedding into HEVC for lightfield image coding. More recently, in (Li et al. 2015), a displacement intra prediction with multiple hypothesis method is proposed for both lightfield image and video content. Please refer to the Chapter 6 for more details on this multiple hypothesis lightfield coding method.

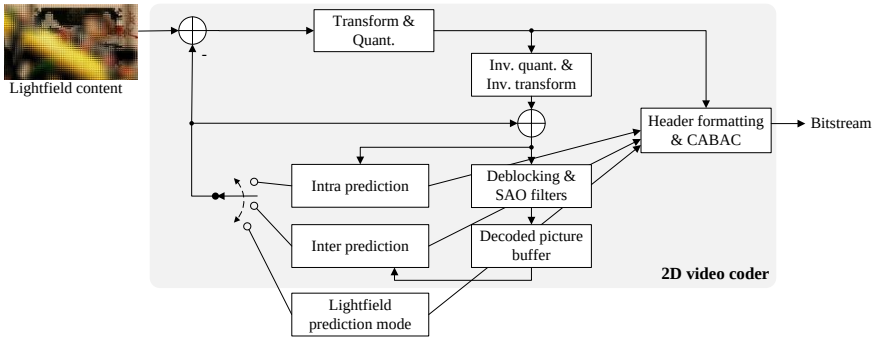


Fig. 12. Basic diagram for a lightfield raw data-based coding approach based on HEVC

The advantage of these coding schemes is that they explore the particular correlation of the lightfield content without requiring any explicit knowledge about the used optical system (e.g., microlens' size, focal length, and distance of the microlenses to the image sensor). Although these parameters may be provided by camera makers, many of them are highly dependent on the manufacturing process, being different from camera to camera. For instance, the fabrication process results in microlenses that may vary slightly in shape, size, and relative position, needing a very careful and complex calibration process in the lightfield camera. For this reason, using compression and rendering tools that are less dependent to these calibration processes would be advantageous for supporting a vaster selection of devices without increasing the complexity.

On the other hand, although these coding schemes achieve significant compression gains when compared to the existing state-of-the-art alternatives, transmitting the entire lightfield data without a scalable bitstream may represent a serious problem since the user needs to wait until the entire content of each picture arrives before it can be visualized, independently of the used type of display and level of interaction the user may want to do with it.

2.5.2. Multiview-based approach

Some coding schemes propose to decompose the lightfield data into several viewpoint sequences to be represented as a multiview video (Adedoyin et al. 2007, Dick et al 2011, Shi et al. 2011) which is then coded with a standard multiview video coder, as illustrated in Fig. . A viewpoint image (a.k.a. sub-image) represents an orthographic projection of the complete captured scene in a particular direction, and can be constructed by simply extracting one pixel with the same relative position from each MI. In (Dick et al 2011, Shi et al. 2011), a coding approach based on the Multiview Video Coding (MVC) (Vetro et al. 2011) extension of H.264 standard is proposed to jointly exploit temporal motion and disparity between adjacent viewpoint images. Therefore, the sequence of each viewpoint is encoded using MVC by defining different scanning orders and coding configurations.

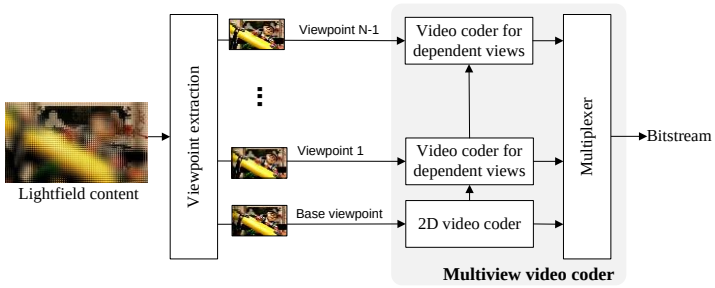


Fig. 13 Basic architecture for a lightfield coding scheme based on a multiview video codec

Although scalability and backward compatibility is guaranteed by using a standard multiview video codec, a drawback of these coding schemes is that they usually consider computer generated sequences with a small number of viewpoint images (up to 9), while this number is typically much higher for natural lightfield content (usually, more than 50). Consequently, these coding schemes become more complex and with a larger amount of header information, when applied to natural content.

Since rendering viewpoint images usually produces very low-resolution images with aliasing (Bishop and Favaro 2009), an alternative to the multiview representation based on these viewpoint images is presented in (Conti et al. 2013), as shown in Fig. 14. In this case, the lightfield content is decimated into 2D views with larger resolution than viewpoint images by using the rendering algorithms proposed in Georgiev and Lumsdaine 2010. Hence, a scalable coding solution is proposed to support backward compatibility with 2D representation (base layer) and also with the current stereo and multiview representation (in one or more enhancement layers). Finally, the top enhancement layer supports the entire lightfield content. For more details about this scalable coding approach, please refer to the Chapter 6.

This scalable coding architecture is able to support a diverse range of consumption environments and devices. On the other hand, the end-user still needs to receive the entire lightfield bitstream to have a viewing experience with the novel and appealing interaction functionalities supported by this type of content (such as changing focus and depth of field).

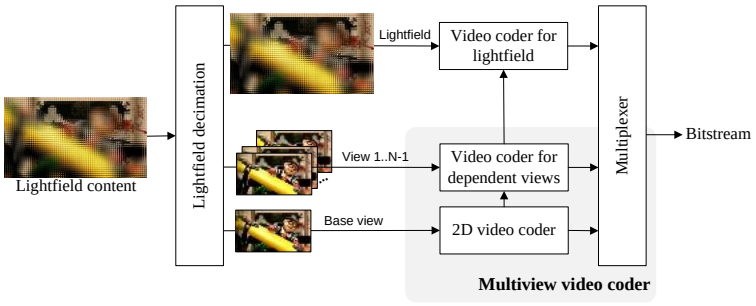


Fig. 14 Basic scalable lightfield coding architecture for backward compatibility with 2D, stereo and multiview representation

2.5.3. Sub-sampled grid of MIs plus disparity approach

Other coding schemes propose to represent the lightfield data by a sub-sampled set of MIs with their associated disparity information (Piao and Yan 2010, Choudhury and Chaudhuri 2014, Graziosi et al. 2015, Sjöström et al. 2015). As firstly proposed in (Piao and Yan 2010), the grid of MIs is sub-sampled to remove the redundancy between neighboring MIs and to achieve compression. Thus, only the remainder subsampled set of MIs and associated disparity data are encoded and transmitted, as depicted in Fig. a. At the decoder side, the lightfield data is reconstructed by simply applying a disparity shift (in (Choudhury and Chaudhuri 2014, Sjöström et al. 2015) or by using a Depth Image Based Rendering (DIBR) algorithm modified to support the multiple MIs as input views (in (Graziosi et al. 2015)), and followed by an inpainting algorithm to fill in the missing areas.

However, in real-world images, the disparity/depth information is estimated from the acquired lightfield raw data, which introduces inaccuracies. Hence, the quality of the reconstructed MIs – and, consequently, the quality of rendered views – is severely affected by these inaccuracies at the encoder side. Additionally, due to occlusion problems and quantization errors when (lossy) encoding this disparity/depth maps, some synthesized MIs might present too many missing areas to be filled, thus introducing even further inaccuracies. The reconstruction artifacts are even more challenging for MI synthesis because of the small angle-of-view (which is intrinsically limited by the pitch of the microlenses).

For this reason, instead of uniformly selecting the MIs, the selection is performed adaptively in (Choudhury and Chaudhuri 2014, Graziosi et al. 2015), so as to obtain better view reconstruction. For this, extra MIs are selected by identifying possible hole-causing regions, increasing considerably the bits consumption. In (Sjöström et al. 2015), the entire lightfield image is also encoded and transmitted in an enhancement layer, as shown in Fig. 14b, so as to provide better rendering views. More details about this coding approach can be seen in the Chapter 6.

The main advantage of incorporating the disparity information into the bitstream is that it facilitates the support of a larger variety of displays and larger levels of user's interaction. However, a common characteristic of the aforementioned approaches is that the quality of rendered views is negatively affected by the inaccuracies in the synthesis of the missing MIs.

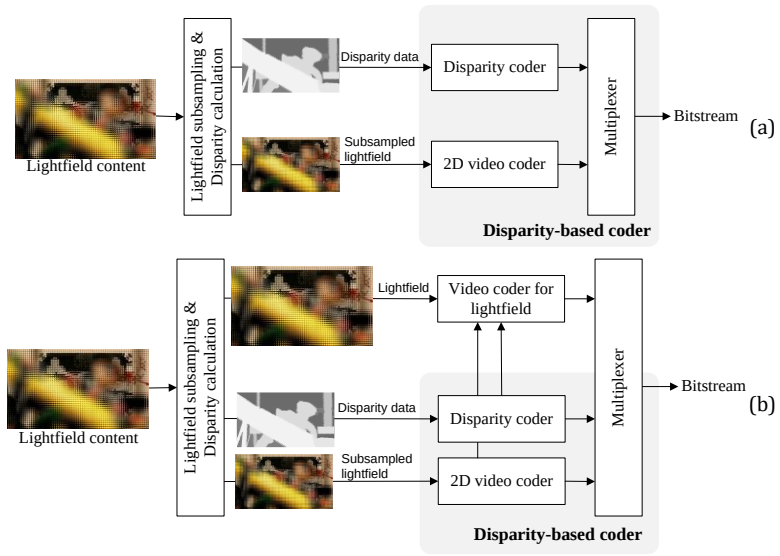


Fig. 15 Basic lightfield coding architectures for sub-sampled grid of MIs plus disparity approach

2.6. Free navigation and free viewpoint television

Free-Viewpoint Television (FTV) is an interactive video service that provides the ability for a viewer to navigate freely around a scene (Tanimoto et al. 2012). Such service is also simply called Virtual navigation or Free Virtual navigation. A viewer watches the scene in an arbitrary direction and from virtual viewpoints on an arbitrary navigation trajectory. At each virtual viewpoint, the corresponding view has to be synthesized and made available at the receiver. Possibly many viewers share the same FTV service, and each viewer navigates independently. View synthesis may use either the distributed model where views are synthesized independently in each receiver, or the centralized model where views requested by all viewers are synthesized in the servers of the service provider. The distributed model requires high transmission bandwidth in server-to-viewer downlinks and significant processing power of viewer terminals. On the other hand, the centralized model suffers from delays in the bidirectional server-to-terminal communications **Error! Reference source not found.**, similarly to networked gaming. Therefore, both models are considered for future applications.

An FTV system requires efficient techniques for multicamera system calibration and video correction, depth estimation and view synthesis as pointed out in previous sections**Error! Reference source not found.**. In a practical FTV system the number of cameras should be limited, and therefore the distances between cameras are large. The cameras are located around a scene, in a roughly circular camera setup (see Fig. 16).

Figure 16. Tripods with wireless camera modules designed and produced at Poznań University of Technology.

Recently, the generic structure of FTV systems has been proposed as shown in Fig.17. Throughout this paper we are going to use this structure that consists of the following functional blocks:

- The video and audio acquisition system,
- The representation server that produces a visual representation of the spatial dynamic scene,
- The rendering servers that serve the requests for synthesis of video and audio at particular virtual locations around a scene,
- The user terminal.

The video and audio acquisition system has to provide data necessary to compute the spatial representation of a scene. Except of video and audio, the data include also some depth information obtained either from pure multiview video analysis or also from depth sensors. The depth acquisition using the depth sensors is conceptually very attractive, but its practical application still faces severe problems related to limited resolutions of the acquired depth maps, limited distance ranges, additional infrared illumination of the scene, synchronization of the video and depth cameras, and sensitivity to the environmental factors including solar illumination. In particular, in this paper we focus on the multiview recording of real events where additional infrared illumination might be unacceptable. Therefore, the considerations in this paper base on the assumption that the depth information is obtained by the video analysis only, and the special depth sensors are not used.

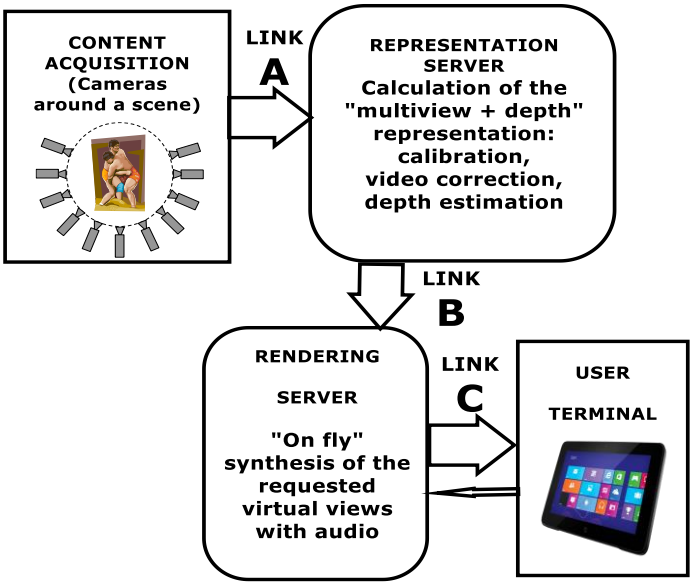


Fig. 17. The general structure of an FTV system - from (Domański et al. 2016)

The video and audio data together with the system calibration data are transmitted via Link A that belongs to the contribution environment, thus needs the high–fidelity compression. As the video data in Link A are yet neither calibrated nor corrected, for video, a standard single-view compression techniques may be used, including both intraframe techniques like M-JPEG 2000 or HEVC All Intra, or interframe studio profiles of AVC or HEVC. Note that simple FTV systems will probably rarely use nonlinear edition as the FTV material does not need any choice of the camera during the production process. The FTV video material does not need camera changes and zooming, as that is done individually by a viewer. If the nonlinear edition is not needed, there is also no need for the random frame access and no need for small error accumulation in the multiple encoding–decoding cycles. Therefore, the requirement to use the intraframe coding may be released, and the standard interframe compression techniques may be used for video. In that way the requested bitrate may be significantly reduced but still the total bitrate will be determined by simulcasting the video streams from multiple cameras plus the audio streams from many microphones.

Acknowledgements

This book chapter was partially supported by COST Action IC1105 - 3D-ConTourNet.

The book chapter was partially supported by National Science Centre, Poland according to the decision DEC-2012/05/B/ST7/01279.

References:

(Aggoun et al. 2013) A. Aggoun, E. Tsekleves, M.R. Swash, D. Zarpalas, A. Dimou, P. Daras, et al., Immersive 3D Holoscopic Video System, IEEE Multimed. 20 (2013) 28–37

(3D Audio 2015) “3D audio”, ISO/IEC International Standard 23008-3: 2015

(Adelson et al. 1991) E. H. Adelson and J. R. Bergen, , M. Landy and J. A. Movshon, Eds.,“The plenoptic function and the elements of early vision,” in Computational Models of Visual Processing. Cambridge, U.K.: MIT Press, 1991, pp. 3–20.

(Aflaki et al. 2014) Aflaki, P.; Hannuksela, M.; Homayouni, M.; Gabbouj, M., “Joint depth and texture filtering targeting MVD compression,” in Proceedings of the 2014 IEEE Visual Communications and Image Processing Conference, pp. 410 – 413, 2014.

(Atzpadin et al. 2004) Atzpadin, N.; Kauff, P.; Schreer, O., "Stereo analysis by hybrid recursive matching for real-time immersive video conferencing", Circuits and Systems for Video Technology, IEEE Transactions on, vol.14, no. 3, pp.321- 334, March 2004.

(AVC 2014) "Advanced video coding", ISO/IEC International Standard 14496-10, 8th Ed., September 2014, and ITU-T Rec. H.264 (V12), 12th Ed., April 2017.

(Balogh and Kovacs 2010) T. Balogh, P. T. Kovács,"Real-time 3D light field transmission", in Proc. Real-Time Image and Video Processing, Proc. SPIE 7724, Brussels, 2010

(Bleyer and Gelautz 2005) M. Bleyer, M. Gelautz, "Graph-based surface reconstruction from stereo pairs using image segmentation", Proceedings of SPIE - The International Society for Optical Engineering, 5665, pp. 288–299, 2005

(Benesty et al. 2008) J. Benesty, J. Chen, and Y. Huang, "Microphone array signal processing", Springer-Verlag, Berlin, 2008.

(Blauert 2013) J. Blauert, Ed., "Technology of binaural listening", Springer-Verlag, Berlin/Heidelberg, 2013.

(Bo et al. 2013) Bo Li; Heng, L.; Koser, K.; Pollefeys, M., "A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern," in Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on , vol., no., pp.1301-1307, 3-7 Nov. 2013. doi: 10.1109/IROS.2013.6696517

(Chen et al. 2008) Chen, Y.; Wang, Y.; Ugur, K.; Hannuksela, M.; Lainema, J.; Gabbouj, M., "The emerging MVC standard for 3D video services," EURASIP Journal on Advances in Signal Processing, vol. 2009, pp. 1–13, Jan. 2008.

(Chen et al. 2013) Chen, Y.; Hannuksela, M.; Suzuki, T.; Hattori, S., "Overview of the MVC+D 3D video coding standard," Journal of Visual Communication and Image Representation, 2013.

(Chen et al. 2016) Ying Chen; Xin Zhao; Li Zhang; Je-Won Kang, "Multiview and 3D Video Compression Using Neighboring Block Based Disparity Vector", IEEE Transactions on Multimedia, Volume: 18, Issue: 4, Pages: 576 – 589, 2016.

(Deng et al. 2012) Deng, H.; Yu, L.; Qui, J.; Zhang, J., "A Joint Texture/Depth Edge-Directed Up-Sampling Algorithm for Depth Map Coding," in Proceedings of the IEEE International Conference on Multimedia and Expo, July 2012.

(Do et al. 2009) Do, L.; Zinger, S.; Morvan, Y.; With, P., "Quality Improving Techniques in DIBR for Free-viewpoint Video," in Proceedings of 3DTV Conference: The True Vision – Capture, Transmission and Display of 3D Video, May 2009.

(Domański et al. 2012) Domański, M.; Grajek, T.; Karwowski, D.; Klimaszewski, K.; Konieczny, J.; Kurc, M.; Łuczak, A.; Ratajczak, R.; Siast, J.; Stankiewicz, O.; Stankowski, J.; Wegner, K.; "New coding technology for 3D video with depth maps as proposed for standardization within MPEG", 19th International Conference on Systems, Signals and Image Processing, IWSSIP 2012, Vienna, Austria, 11-13 April 2012, pp. 401-404.

(Domański et al. 2012a) Marek Domański, Tomasz Grajek, Damian Karwowski, Jacek Konieczny, Maciej Kurc, Adam Łuczak, Robert Ratajczak, Jakub Siast, Jakub Stankowski, Krzysztof Wegner, "Coding of multiple video+depth using HEVC technology and reduced representations of side views and depth maps," 29th Picture Coding Symposium, PCS 2012, Kraków, May 2012, pp. 5-8.

(Domański et al. 2013) Marek Domański, Olgierd Stankiewicz, Krzysztof Wegner, Maciej Kurc, Jacek Konieczny, Jakub Siast, Jakub Stankowski, Robert Ratajczak, Tomasz 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.

(Domański 2015) Marek Domański, "Approximate video bitrate estimation for television services", ISO/IEC JTC1/SC29/WG11 Doc. MPEG M3671, Warsaw, June 2015.

(Domański et al. 2015) M. Domański, A. Dziembowski, D. Mieloch, A. Łuczak, O. Stankiewicz and K. Wegner, "A practical approach to acquisition and processing of free viewpoint video," in 2015 Picture Coding Symposium (PCS), Cairns, QLD, 2015, pp. 10-14.

(Domański et al. 2015a) Marek Domański, Adrian Dziembowski, Tomasz. Grajek, Adam Grzelka, Łukasz Kowalski, Maciej. Kurc, Adam Łuczak, Dawid Mieloch, Robert Ratajczak, Jarosław Samelak, Olgierd Stankiewicz, Jakub Stankowski, Krzysztof Wegner, "Methods of high efficiency compression for transmission of spatial representation of motion scenes", IEEE Int. Conf. Multimedia and Expo Workshops, Torino 2015.

(Domański et al. 2016) M. Domański, M. Bartkowiak, A. Dziembowski, T. Grajek, A. Grzelka, A. Łuczak, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, Krzysztof Wegner, "New results in free-viewpoint television systems for horizontal virtual navigation," in 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, 2016, pp. 1-6.

(Domański et al. 2016a) Marek Domański, Adrian Dziembowski, Adam Grzelka, Dawid Mieloch, "Optimization of camera positions for free-navigation applications", International Conference on Signals and Electronic Systems, ICSES 2016, Kraków, Poland, September 5-7 2016,

(Domański et al. 2017) M. Domański, O. Stankiewicz, K. Wegner, T. Grajek, "Immersive visual media – MPEG-I: 360 video, virtual navigation and beyond", Int. Conference on Systems, Signal and Image Processing, Poznań, May 2017.

(Dziembowski et al. 2016) Adrian Dziembowski, Adam Grzelka, Dawid Mieloch, Olgierd Stankiewicz, Krzysztof Wegner, Marek Domański, "Multiview Synthesis – improved view synthesis for virtual navigation", 32nd Picture Coding Symposium, PCS 2016, Nuremberg, Germany, December 4-7, 2016,

(EBU 2017) EBU Technical Report TR 039, "Opportunities and challenges for public service media in vr, ar and mr", Geneva, April 2017.

(Gao et al. 2016) Yu Gao, Gene Cheung, Thomas Maugey, Pascal Frossard, Jie Liang, "Encoder-Driven Inpainting Strategy in Multiview Video Compression", IEEE Transactions on Image Processing, Volume: 25, 2016, Pages: 134 – 149.

(Gokturk et al. 2004) Gokturk, S.; Yalcin, H.; Bamji, C., "A time-of-flight depth sensor – system description, issues and solutions," in Proc. of the IEEE Conf. Comput. Vision and Pattern Recognition Workshop, Jun. 2004.

(Grand Front Osaka 2013) "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.

(Graziosi et al. 2010) Graziosi, D.; Rodrigues, N.; Pagliari, C.; Faria, S.; Silva, E.; Carvalho, M., "Compressing depth maps using multiscale recurrent pattern image coding," in Electronics Letters, vol.46, no.5, pp. 340-341, March 2010.

(Hannuksela et al. 2013) Miska M. Hannuksela; Dmytro Rusanovskyy; Wenyi Su; Lulu Chen; Ri Li; Payman Aflaki; Deyan Lan; Michal Joachimiak; Houqiang Li; Moncef Gabbouj, "Multiview-Video-Plus-Depth Coding Based on the Advanced Video Coding Standard", IEEE Transactions on Image Processing, Volume: 22, Issue: 9, 2013, Pages: 3449 – 3458.

(Hartley and Zisserman 2015) R. Hartley, A. Zisserman, "Multiple view geometry in computer vision" (2nd ed.), Cambridge Univ. Press, 2015

(Haskell et al.1996) Barry G. Haskell, Atul Puri, Arun N. Netravali, "Digital video: an introduction to MPEG-2", Chapman & Hall, New York, 1996.

(Herre et al. 2015) 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.

(HEVC 2016) "High Efficiency Video Coding", ISO/IEC IS 23008-2, 3rd Ed., October 2017, and ITU-T Rec. H.265, 4th Ed., December 2016.

(Holografika 2017) Holografika,"HoloVizio C80 3D cinema system", Budapest, http://www.holografika.com/Products/NEW-HoloVizio-

C80. html, retrieved on April 21, 2017.

(Hong and Chen 2004) L. Hong, G. Chen, “Segment-based stereo matching using graph cuts”, 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 74-81, 2004

(Isgro et al. 2004) 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.

(Ishida and Shibata 2010) 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.

(ISO 2010) “Report on Experimental Framework for 3D Video Coding,” ISO/IEC JTC1/SC29/WG11, N11631, October 2010.

(ISO 2015) ”Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation – update”, ISO/IEC JTC1/SC29/WG11 Doc. N15733, October 2015, Geneva, Switzerland

(ISO 2017) "Requirements for Omnidirectional Media Format" ISO/IEC JTC1/SC29/WG11 Doc. N 16773, April 2017, Hobart, Australia.

(ISO 2017a) “Call for Proposals for Point Cloud Coding V2”, ISO/IEC/JTC1/SC29/WG11, Doc. N16763, April 2017, Hobart, Australia.

(Jones et al. 2007) A. Jones, I. McDowall, H. Yamada, M. Bolas, P. Debevec, “Rendering for an interactive 360° light field display,” ACM Trans. Graphics, vol. 26, no. 3, art. 40, Jul 2007

(Jorissen et al. 2015) L. Jorissen, P. Goorts, S. Rogmans, G. Lafruit, P. Bekaert, “Multi-camera epipolar plane image feature detection for robust view synthesis”, 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2015

[Kang and Ho 2010) Y.S. Kang, Y.S. Ho, “High-quality multi-view depth generation using multiple color and depth cameras”, IEEE International Conference on Multimedia and Expo 2010, pp. 1405-1410, 2010

(Kim et al. 2013) C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, M. Gross, “Scene reconstruction from high spatio-angular resolution light fields,” ACM Trans. Graph, vol. 32, no. 4, art. 73, Jul 2013

(Lafruit et al. 2016) G. Lafruit, M. Domański, K. Wegner, T. Grajek, T. Senoh, J. Jung, P. Kovács, P. Goorts, L. Jorissen, A. Munteanu, B. Ceulemans, P. Carballeira, S. García, M. Tanimoto, “New visual coding exploration in MPEG: Super-MultiView and Free Navigation in Free viewpoint TV,” in IST Electronic Imaging, Stereoscopic Displays and Applications XXVII, San Francisco 2016, pp. 1-9.

(Lee and Ho 2010) Lee, S.; Ho, Y., “View-consistent multiview depth

estimation for three-dimensional video generation", 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), pp.1-4, June 2010.

(Lei et al. 2015) Lei, J.; Zhang, C.; Fang, Y.; Gu, Z.; Ling, N.; Hou, C., "Depth Sensation Enhancement for Multiple Virtual View Rendering," IEEE Transactions on Multimedia, vol. 17, no. 4, pp. 457-469, April 2015.

(Lucas et al. 2015) Lucas, L.; Wegner, K.; Rodrigues, N.; Pagliari, C.; Silva, E.; Faria, S., "Intra Predictive Depth Map Coding using Flexible Block Partitioning," IEEE Transactions on Image Processing, vol.24, no.11, pp.4055-4068, November 2015.

(Merkle et al. 2015) Merkle, P.; Muller, K.; Marpe, D.; Wiegand, T., "Depth Intra Coding for 3D Video based on Geometric Primitives," IEEE Transactions on Circuits and Systems for Video Technology, 2015.

(Merkle et al.2012) 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.

(Mieloch et al. 2017) Dawid Mieloch, Adrian Dziembowski, Adam Grzelka, Olgierd Stankiewicz, Marek Domański, "Graph-based multiview depth estimation using segmentation", IEEE International Conference on Multimedia and Expo ICME 2017, Hong Kong, 10-14 July 2017,

(Miller et al. 2006) G. Miller, J. Starck, A. Hilton, "Projective surface refinement for free-viewpoint video," 3rd European Conf. Visual Media Production, CVMP 2006, pp.153-162.

(Min et al. 2010) Min, D.; Yea, S.; Vetro, A., "Temporally consistent stereo matching using coherence function", 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), pp.1-4, June 2010.

(Montserrat et al. 2009) T. Montserrat, J. Civit, O. Escoda, J. Landabaso, "Depth estimation based on multiview matching with depth/color segmentation and memory efficient belief propagation", IEEE International Conference on Image Processing, pp. 2329-2332, 2009

(Mori et al. 2008) Mori, Y.; Fukushima, N.; Fujii, N.; Tanimoto, M., "View Generation with 3D Warping using Depth Information for FTV," in Proceedings of 3DTV Conference: The True Vision – Capture, Transmission and Display of 3D Video, May 2008.

(MPEG-2 2012) "Generic coding of moving pictures and associated audio information: Video", ISO/IEC Int. Standard 13818-2: 2013 and ITU-T Rec. H.262 (V3.1),2012.

(MPEG Surround 2007) ISO/IEC IS 23003-1: 2007, "MPEG audio technologies -- Part 1: MPEG Surround".

(Müller et al. 2011) Müller K., Merkle P., and Wiegand T., "3D Video Representation Using Depth Maps", Proc. IEEE, vol. 99, no. 4, pp. 643–656, Apr. 2011

(Müller et al. 2013) Karsten Müller, Heiko Schwarz, Detlev Marpe, Christian Bartnik, Sebastian Bosse, Heribert Brust, Tobias Hinz, Haricharan Lakshman, Philipp Merkle, Franz Hunn Rhee, Gerhard Tech, Martin Winken, Thomas Wiegand, "3D High-Efficiency Video Coding for Multi-View Video and Depth Data", IEEE Transactions on Image Processing, Volume: 22, Issue: 9, 2013, Pages: 3366 – 3378.

(NICT 2011) NICT News, Special Issue on Stereoscopic Images, no. 419, November 2011.

(Oh et al. 2010) Oh, K.; Yea, S.; Vetro, A.; Ho, Y., "Virtual View Synthesis Method and Self-Evaluation Metrics for Free Viewpoint Television and 3D Video," International Journal of Imaging Systems and Technology, vol. 20, no. 4, pp. 378-390, December 2010.

(OMF 2017) "Omnidirectional Media Format", ISO/IEC DIS 23090-2, Doc. ISO/IEC JTC1/SC29/WG11 N16824 April 2017, Hobart, Australia

(Sandberg et al. 2011) D. Sandberg, P. E. Forssen and J. Ogniewski, "Model-based video coding using colour and depth cameras," in 2011 International Conference on Digital Image Computing: Techniques and Applications, Noosa, QLD, 2011, pp. 158-163

(Samelak et al. 2016) Jarosław Samelak, Jakub Stankowski, Marek Domański, "Adaptation of the 3D-HEVC coding tools to arbitrary locations of cameras", International Conference on Signals and Electronic Systems, Kraków, 2016.

(Samelak et al. 2017) Jarosław Samelak, Olgierd Stankiewicz, Marek Domański, „Do we need multiview profiles for future video coding generations ?, Doc. ISO/IEC JTC1/SC29/WG11 M41499 October 2017, Macau, China

(SAOC 2016) "Spatial Audio Object Coding (SAOC)", ISO/IEC IS 23003-2: 2016, 2nd Ed.

(Sen et al. 2013) X. Sen, Y. Li, L. Qiong, X. Zixiang, "A gradient-based approach for interference cancelation in systems with multiple Kinect cameras", 2013 IEEE International Symposium on Circuits and Systems, pp. 13-16, 2013

(Shao et al. 2016) Feng Shao; Weisi Lin; Gangyi Jiang; Mei Yu, "Low-Complexity Depth Coding by Depth Sensitivity Aware Rate-Distortion Optimization", IEEE Transactions on Broadcasting, Volume 62, Issue 1, pp. 94 – 102, 2016.

(Smolic et al. 2005) A. Smolic, et al., "3D video objects for interactive applications." European Signal Proc. Conf. EUSIPCO 2005.

(Stamos and Allen 2000). Stamos and P. K. Allen, "Integration of range

and image sensing for photo-realistic 3D modeling," in Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings, San Francisco, CA, 2000, pp. 1435-1440 vol.2

(Stankowski et al. 2015) Jakub Stankowski, Łukasz Kowalski, Jarosław Samelak, Marek Domański, Tomasz Grajek, Krzysztof 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, 8-10 July 2015.

(Stankiewicz et al. 2013) O. Stankiewicz, K. Wegner, M. Tanimoto, M. Domański, "Enhanced Depth Estimation Reference Software (DERS) for Free-viewpoint Television", ISO/IEC JTC1/SC29/WG11 Doc. MPEG M31518, Geneva, 2013

(Stankiewicz et al. 2015) Olgierd Stankiewicz, Marek Domański, Krzysztof Wegner, "Estimation of Temporally-Consistent Depth Maps from Video with Reduced Noise", 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 3DTV- Con 2015, Lisbon, Portugal, 8-10 July 2015,

(Sun et al. 2003) J. Sun, N.N. Zheng, H.Y. Shum, "Stereo Matching Using Belief Propagation," IEEE Transaction on Pattern Analysis and Machine Intelligence, 25(7), pp. 787-800, 2003

(Tanimoto 2006) M. Tanimoto, "Overview of free viewpoint television", Signal Proc.: Image Communic., vol. 21, 2006, pp. 454-461.

(Tanimoto 2011) Tanimoto, M.; Tehrani, M.; Fujii, T.; Yendo, T., "Free-viewpoint TV – A Review of the Ultimate 3DTV and its Related Technologies," IEEE Signal Processing Magazine, pp. 67-76, January 2011.

(Tanimoto et al. 2005) M. Tanimoto, T. Fujii, T. Senoh, T. Aoki, Y. Sugihara, "Test Sequences with Different Camera Arrangements for Call for Proposals on Multiview Video Coding," ISO/IEC JTC1/SC29/WG11/M12338, Poznan, 2005

(Tanimoto et al. 2012) M. Tanimoto, M. Panahpour, T. Fujii, T. Yendo, "FTV for 3-D spatial communication," Proceedings of the IEEE, vol. 100, Issue 4, pp. 905-917, Feb. 2012.

(Tao et al. 2009) Tao, S.; Chen, Y.; Hannuksela, M.; Wang, Y.; Gabbouj, M.; Li, H., "Joint Texture and Depth Map Video Coding Based on the Scalable Extension of H.264/AVC," in Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 2353-2356, May 2009.

(Tech et al. 2016) Gerhard Tech, Ying Chen, Karsten Müller Jens-Rainer Ohm, Anthony Vetro, Ye-Kui 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.

(Tran and Harada 2013) Tran, A.; Harada, K., "View Synthesis with Depth Information based on Graph Cuts for FTV," in Proceedings of the 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision, pp. 289-294, February 2013.

(USAC 2016) "Unified Speech and Audio Coding (USAC)", ISO/IEC IS 23003-2: 2016 (2nd Ed.)

(Vetro et al. 2011) Anthony Vetro, Thomas Wiegand, Gary 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.

(Wang 2015) Q. Wang, "Computational models for multiview dense depth maps of dynamic scene," 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015

(Wegner et al. 2013) Wegner, K.; Stankiewicz, O.; Tanimoto, M.; Domanski, M.; "Enhanced View Synthesis Reference Software (VSRS) for Free-viewpoint Television," ISO/IEC JTC1/SC29/WG11 MPEG2013/M31520 October 2013, Geneva, Switzerland.

(Wegner et al. 2014) Wegner, K.; Stankiewicz, O.; Domański, M.; "Fast View Synthesis using platelet-based depth representation," 21th International Conference on Systems, Signals and Image Processing, IWSSIP 2014, Dubrovnik, Croatia, May 2014.

(Wei et al. 2013) 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.

(WIKI 2017) https://en.wikipedia.org/wiki/360-degree_video, as Octiber 28th, 2017.

(Wilburn et al. 2005) B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, M. Levoy, "High performance imaging using large camera arrays" ACM Trans. Graphics, vol. 24, no. 3. pp 765-776, July 2005

(Xu et al. 2014) Xu, J.; Yan, F.; Cao, X., "Stereoacuity-guided Depth Image based Rendering," in Proceedings of the IEEE International Conference on Multimedia and Expo, July 2014.

(Yang et al. 2011) Yang, X.; Lui, J.; Sun, J.; Li, X.; Liu, W.; Gao, Y., "DIBR based View Synthesis for Free-viewpoint Television," in Proceedings of 3DTV Conference: The True Vision – Capture, Transmission and Display of 3D Video, May 2011.

(Zarb and Debono 2014) Zarb, T.; Debono, C., "Depth-based Image Processing for 3D Video Rendering Applications," in Proceedings of the 21st International Conference on Systems, Signals and Image Processing, pp. 215-218, May 2014.

(Zhang 2000) Z. Zhang, "A Flexible New Technique for Camera

Calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330-1334, 2000

(Zhang et al. 2010) Zhang, J.; Hannuksela, M.; Li, H., "Joint Multiview Video Plus Depth Coding," in Proceedings of the 2010 IEEE 17th International Conference on Image Processing, September 2010.

(Ziegler et al. 2014) 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.

(Zilly et al. 2015) F. Zilly, M. Schoberl, M. Ziegler, J. Keinert, S. Foessel, "Light-Field Acquisition System That Facilitates Camera and Depth-of-Field Compositing in Post-Production," SMPTE Motion Imaging Journal, vol. 124, no. 1, pp. 16-21, Feb 2015