



HAL
open science

Introduction to Immersive Video Technologies

Martin Alain, Emin Zerman, Cagri Ozcinar, Giuseppe Valenzise

► **To cite this version:**

Martin Alain, Emin Zerman, Cagri Ozcinar, Giuseppe Valenzise. Introduction to Immersive Video Technologies. Immersive Video Technologies, Academic Press, 2022, 9780323917551. 10.1016/b978-0-32-391755-1.00007-9 . hal-03831371

HAL Id: hal-03831371

<https://hal.science/hal-03831371>

Submitted on 7 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Introduction to Immersive Video Technologies

1

Martin Alain^{*,a}, Emin Zerman^{*,b}, Cagri Ozcinar^c, and Giuseppe Valenzise^d

^aHuawei Ireland Research Centre, Dublin, Ireland, ^bSTC Research Center, Mid Sweden University, Sundsvall, Sweden, ^cSamsung UK, United Kingdom, ^dCentral-Supelec, France

Chapter Points

- What is immersion
- Degrees of freedom
- The plenoptic function
- Immersive imaging modalities
- Immersive video content delivery pipeline

ABSTRACT

Immersive imaging technologies have become a topic of great interest in recent years due to the convergence of maturing research from different fields and broad aspects of signal and image processing, including but not limited to: computer vision, computer graphics, computational imaging, optics, and recent advances in deep learning. In particular, recent research and developments in these areas helped achieve better or faster content creation and image/video processing. This enables the design of complete practical systems covering capture to display for immersive imaging. In this chapter, we provide the theoretical background for the rest of the book and introduce key concepts commonly used for traditional imaging systems and immersive video technologies. We also describe the stages of the immersive imaging technologies from content capture to display and quality assessment for three immersive imaging technologies: omnidirectional video, light fields, and volumetric (also known as free-viewpoint) video.¹

KEYWORDS

Immersive imaging, Plenoptic function, Omnidirectional video, Light field, Volumetric video, Point cloud, Textured mesh, Content delivery pipeline

1.1 Introduction

Imaging technologies enable humankind to capture and store the visual information from real world scenes. Although traditionally images have been stored on physical

¹*Martin Alain and Emin Zerman were with V-SENSE, Trinity College Dublin, Dublin, Ireland at the time of writing this chapter.

media (e.g., photographic film), with the advent of digital image processing, we can capture, store, compress, and transmit images and videos digitally and in real time. This enabled *telepresence* by delivering the visual information to distant locations. The telepresence term refers to the phenomenon that a human operator develops a “sense of being physically present at a remote location through interaction with the system’s human interface” [1]. Following that, the term *presence* is also coined for “being there” for other virtual environments [2], as well as *immersion*, for “concentration to the virtual environment instead of real world” [3]. Immersion is considered as one of the factors which are necessary for presence [4]. Therefore, the technologies which try to provide a virtual presence are called *immersive* imaging technologies.

The state-of-the-art immersive video technologies extend the visual sensation, augment the viewer “presence”, and provide the viewer with a higher degree of freedom (DoF) than what traditional displays offer (see Figure 1.1). The traditional imaging systems record the scene from only a single viewpoint selected by the content creator, which provides essentially zero DoF, as the viewer does not have any freedom over the viewpoint selection. Instead, the immersive imaging systems can provide more than three DoF. 3DoF generally refers to the rotational movement (i.e., changes in yaw, roll, pitch angles), while 6DoF refers to both rotational and translational (e.g., changes in the location in 3D space) movement. 3DoF+ denotes a system that enables limited spatial movement in addition to unrestricted 3DoF rotational movement. Different imaging modalities make use of one or more of these DoF categories.

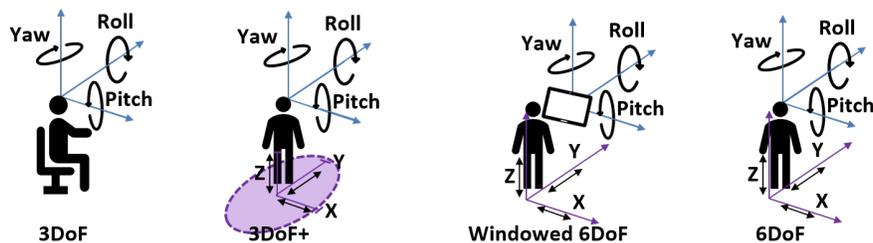


FIGURE 1.1

Increasing degrees of freedom (DoF) provide a more immersive experience.

Recently, these imaging technologies gained a lot of interest in academia and industry, including standardization bodies (e.g., JPEG [5] and MPEG [6,7]), and it became a hot research topic. In this book chapter, we introduce the main concepts underlying the two main axes of this book for immersive video technologies: three different imaging modalities and the stage in the content delivery pipeline. We also discuss the current challenges for immersive video technologies. Please note that we focus in this chapter and this book in general on imaging systems aiming at capturing and reproducing real world scene rather than computer generated content.

1.2 What is immersion?

This book focuses on how we capture, process, transmit, use, and perceive various immersive video technologies. However, before advancing any further, we should answer the first question our readers might have: “*What is immersion?*”. One can also ask another question in a different format: “*What makes a video immersive?*”.

To answer this question, we need to understand how immersion and other related terms are defined in the scientific literature by the cognition and virtual reality experts. The following subsection focuses on description and discussion of these definitions. We then define what immersion means in the context of immersive video, and we discuss which aspects are important to *make a video immersive*. In the next subsection, we also discuss how immersive video technologies relate to extended reality concepts.

1.2.1 Definitions

The concepts of presence and immersion are discussed in great detail by many scholars, including the ones working on human cognition, robotics, and virtual reality. In this section, we will describe the mainly used four terms that are relevant for the scope of this book: telepresence, presence, embodiment, and immersion.

Telepresence: *Telepresence* is the first term to be coined that is relevant to presence and other relevant terms. The term was coined as a response to the needs of the robotics community. The term describes the relationship between the human operator and the environment in which a remote machine is located, where the human operator would get a “sense of being physically present at a remote location through interaction with the system’s human interface” [1]. The term is then adopted by the virtual reality community as well.

Presence: The term *presence* was initially referring to the “experience in natural surroundings” while telepresence was used for the experience in mediated environments [8]. In time, this distinction based on mediation of the environment was viewed unnecessary [2] and *presence* term started to be used for both natural and mediated environments. It is defined as “being there” [8] or “perceptual illusion of nonmediation” [4] (i.e., as if the virtual environment was “real”). As one of the loaded terms among others, presence can have many different lenses to look from. Lombard & Ditton [4] identified six different viewpoints and aspects that define and affect presence. Immersion is described as one of these aspects. Presence is considered an essential element for immersive technologies [9].

Embodiment: Since it is used in virtual reality terminology, the term *embodiment* seems related to presence and immersion concepts. Nevertheless, the term embodiment relates to an avatar which represents the users’ body in a virtual environment. This avatar can be either a photorealistic one or not. If the users of a virtual environment (or a virtual reality application) “embodies” the virtual avatar they are given, they can use the body for physical and social needs in the said virtual environment and do not experience any discomfort during their activities [10]. Currently, most of the applications for the considered immersive video modalities do not employ a virtual

avatar, and there is little need to consider embodiment in most of the applications.

Immersion: The term *immersion* comes from the English word “immerse” meaning “to become completely involved in something” or “to put something or someone completely under the surface of a liquid” [11]. One of the earliest description in the virtual reality literature describes it as “a term that refers to the degree to which a virtual environment submerges the perceptual system of the user in computer-generated stimuli” [8]. This is similar to covering over (or blocking) the users’ senses (both physical and psychological – discussed more in the next subsection) with the virtual environment or virtual reality application. It is also mentioned that immersion can be measured; physical immersion can be measured by identifying number of senses that are covered by the environment and psychological immersion needs to be reported by the user [4].

1.2.2 Immersion

Immersion can be perceptual (i.e., physical or sensory) or psychological [4]. The sensory immersion is achieved by “shutting off” as many senses as possible, including sight (with head-mounted displays), hearing (with audio), touch (with haptics), smell and taste (with olfactory devices). On the other hand, psychological immersion can happen on the cognition level if the user is involved with the material enough and feels lost in the environment. Since immersion has both physical and psychological aspects, any number of interesting activities can achieve user immersion, including daydreaming, reading a book, and cinema (i.e., traditional video). Nevertheless, in this book, we only consider the video technologies that attempt at both physical and psychological immersion. That is, the users should see a different (or augmented) visual and feel present in the prepared environment.

We identified two aspects that are crucial in defining the immersive video technologies: realism and interactivity.

It is hard to define realism in one way since it is our brains that define what “real” is. Lombard & Ditton argue that there are two types of realism: social realism and perceptual realism [4]. Social realism is the type of realism that focuses on the conceptual relationships in the environment, especially between people or agents. Therefore, a video game can be compelling or feel “real” to players even though the avatars or objects in the game do not look as they do in real life. Perceptual realism, on the other hand, focuses on recreating the actual 3D world with highest fidelity, and the perceptual realism for visuals is sometimes called photorealism. Since the users can believe the environment is “real” with social realism, we understand and acknowledge that photorealism (or perceptual realism) is not necessary for immersion in video technologies. Nevertheless, since realism is a very important part of video technologies, the immersive video modalities in this book put a much heavier focus on perceptual realism.

The second aspect, interactivity, is also key in how video technologies become immersive. It is argued that interaction is a crucial element in perceiving a technological system as a social agent instead of hardware [4], which can be very important

in achieving immersion. Recent immersive imaging systems enable and promote interaction in a far greater degree than traditional video, and this is also supported by new lightweight wearables, haptics controllers, and headsets. Furthermore, the low latency between users, which come with increased connection speeds, promotes much smoother user-to-user interaction.

1.2.3 Extended reality

With the help of advanced imaging and display technologies, humankind now can create alternative distinct realities, different from the actual real life we are living in. The early developments for “virtual” realities began after the displays could be made small enough so that an individual can use it to create a different environment. The concept of “suspension of disbelief” was already known to humankind for centuries in literary works such as novels or drama. That is, the audience deliberately chooses to forget that the literary piece they are experiencing is not real and is fiction. The term was coined by an English poet, Samuel Taylor Coleridge, in 1817 [12], and its use (or re-use) in computer graphics and imaging systems were only possible after the medium was ready to create such virtual environments.

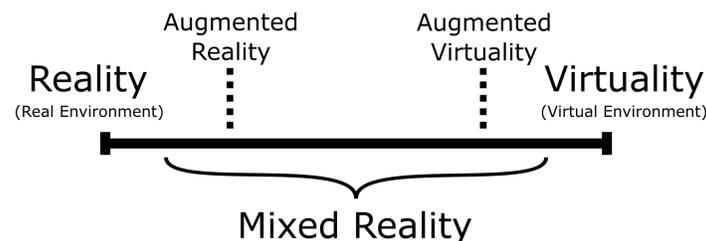


FIGURE 1.2

The reality-virtuality continuum suggested by Milgram et al. [13].

In 1995, Milgram et al. [13] suggested that the changes in reality can be shown on a spectrum (or a continuum), as shown on Figure 1.2. On one end, our actual real environment is situated, and on the other hand the virtual environment is situated. With small changes and additions, the relationship between reality and virtuality can be changed. Adding virtual elements to reality, we obtain “*augmented reality*” systems, which augments (or improves) our reality with virtual elements. Adding real elements to virtuality, we can obtain “*augmented virtuality*” systems, which amends the virtual environment with elements from real world. Everything in between was described as mixed reality, as the systems lying on this part of the spectrum can have elements from both reality and virtuality.

The virtual reality term was popularized around the 1980s even though the first attempts for virtual reality precede that time. Since the aim of creating the first virtual reality systems was to create an alternative reality to our own, the concept of presence

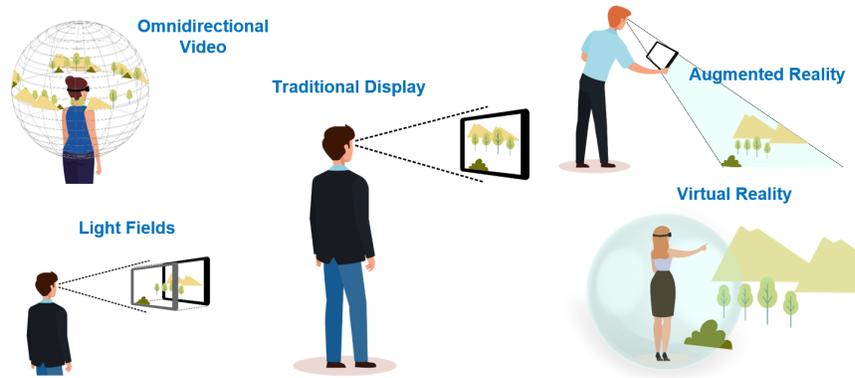


FIGURE 1.3

The immersive imaging modalities provide additional degrees of freedom compared to traditional display.

was very important. So, the virtual reality community continued on developing the terminology that is discussed above. VR mainly focuses on creating new virtual environments for users to feel “presence” in by cutting their senses off the actual world.

Augmented reality, on the other hand, aims to build on top of the actual world and light from the real world to augment (or enhance) our experience in reality. These enhancements can be any form of modality, and for video technologies, it generally includes either a computer generated imagery, text, or shapes with or without colors.

Extended reality is generally used as an umbrella term that refers to whole continuity spectrum (including augmented, virtual, and mixed realities) used in conjunction with equipment that allows capture & display of and interaction with the aspects of the said realities. The immersive video technologies can sometimes cover the real world (e.g., in the case of omnidirectional video – see Part II), can be used to augment the reality with volumetric 3D media (e.g., volumetric video – see Part IV), or can be versatile and be used for all extended reality applications (e.g., light fields – see Part III).

Metaverse is also a popular concept and term that attracts a lot of attention nowadays. Although there are different viewpoints and definitions, one definition of metaverse can be considered as “an immersive Internet as a gigantic, unified, persistent, and shared realm” [14]. As this concept aims to bring all digital assets under a united umbrella, immersive video technologies will be important as much as other extended reality technologies. Nevertheless, metaverse is not discussed further in the book.

In the next section, we discuss how immersion is used for video technologies and different imaging modalities covered in this book.

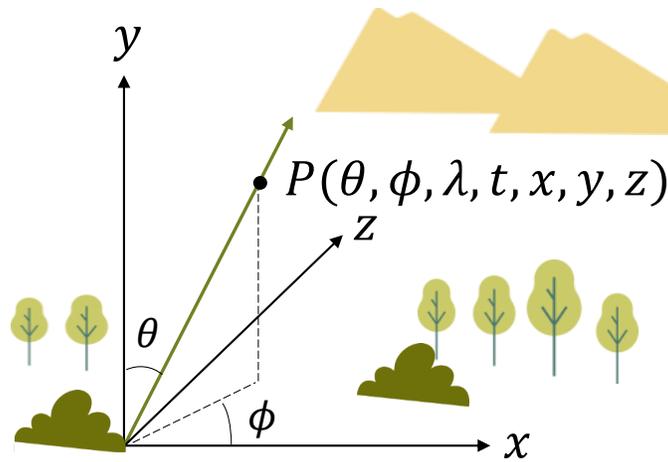


FIGURE 1.4

The 7D plenoptic function measures the intensity of light seen from: any point in space (x, y, z) , any angular viewing direction (θ, ϕ) , over time (t) , for each wavelength (λ) .

1.3 Immersive video

In this section, we first start with the description of a foundational technical concept: the plenoptic function, which describes the light rays in space. We then briefly mention the historical perspective and evaluation of how video has been becoming immersive. Finally, we describe which imaging modalities are discussed within the scope of this book.

1.3.1 Foundations: The plenoptic function

Traditional imaging technologies focus on projecting the 3D world onto a 2D plane, and they are designed to acquire and display visual media from a fixed viewpoint. Immersive video technologies, on the other hand, aim to allow the user to immerse themselves in the presented visual media by providing a more thorough reconstruction of the 3D world. The new immersive technologies vastly change the view-scape compared to traditional video displays, as illustrated in Figure 1.3.

To formulate the acquisition processes, we can consider light as a field and assume that the 7-dimensional plenoptic function $P(\theta, \phi, \lambda, t, x, y, z)$ describes all possible visible light [15], considering every location x, y, z , every angle θ, ϕ , each wavelength λ , and each time instant t , as shown in Figure 1.4. The plenoptic function remains a theoretical concept, which immersive imaging modalities aim to replicate or approximate, based on some assumptions, for example, the radiance of a light ray remains constant along its path through empty space. While it remains a theoretical concept,

its study has led to the development of the broad Image Based Rendering (IBR) field, which goal is to capture images of the real world such that they can then be used to render novel images corresponding to different viewpoints, hence achieving 6DoF in a virtual reproduction of a real world scene. Note that taking into account the viewing direction allows capturing and reproducing subtle lighting effects such as reflections and non-specularities. As discussed above, increasing the DoF is a key feature of immersive imaging systems. Thus, to some extent, we can consider that all imaging modalities discussed are subsampling the plenoptic function.

1.3.2 Historical perspective and evolution

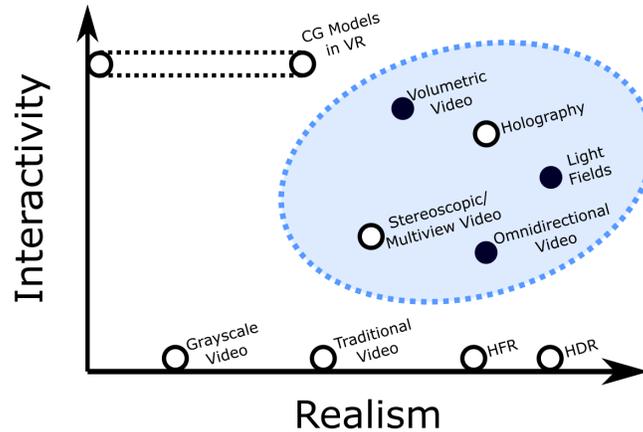
During the evolution of digital imaging technologies, how images and video are represented always changed. This started with improving the acquisition and display frame rate at first for the black-and-white (or rather grayscale) images. Following this, introduction of sound to videos and introduction of color always increased the power of video on “suspending the disbelief”.

In our search to increase the immersiveness, stereo 3D was developed, which created an intense hype during the 1950s. At first as anaglyph projection systems (which is not ideal for color representations) and light polarization were used at cinemas. Around the 1970s, the active shutter systems were developed. Advances in display technologies in the 1980s and 1990s also made glasses-free 3D perception by a technique called autostereoscopy possible in consumer electronics and household TV sets.

It can be noticed that public’s interest in these technologies follow a wave pattern that is similar to the Gartner hype cycle. Whenever there is a new technology, a very prominent expectation wave comes. However, this wave then fades until the technology advances enough to produce capable devices to realize what was imagined. Then, this creates a new expectation wave for more advanced technologies. The above history of stereoscopic 3D shows an example of this. Similarly, the first virtual reality headsets were developed in the 1980s and virtual reality (or at least the idea of alternative/virtual realities) became popular. Until recently, virtual reality was not prominent in the public eye since there were no equipment affordable by the consumers. With the new affordable headsets, VR has become popular again as people can now buy and keep a headset at their homes.

Similarly, other immersive technologies came to fruition recently. Light fields were first discussed in the 1990s, but until now there were not many applications possible due to physical and computational limitations. Similarly the idea of volumetric video was around since the 70s which was popularized by R2D2’s projection of Princess Leia on Star Wars. Recent advances in image acquisition, processing capabilities, deep learning techniques, and display technologies make realization of immersive video technologies possible for 3DoF, 3DoF+, and 6DoF applications.

On the other hand, computer graphics has been exploring virtual visual content creation and 6DoF rendering since its beginning. This was possible thanks to having no physical limitations that forbade capturing light (or simulated light) going towards

**FIGURE 1.5**

A comparison of interactivity and realism aspects of various modalities of video technologies.

any direction. While originally designed to represent computer generated objects and scene, typically using polygons to represent the geometry and texture map images to represent the appearance, the tools developed in the computer graphics field can be useful for representing immersive imaging content. Indeed, such tools have been optimized for years for efficiency and visual quality. Once the scene geometry and objects are known, efficient rendering algorithms can be used (which include hardware acceleration) to render images corresponding to the desired viewport direction and orientation. While particularly relevant for volumetric videos which relies on geometric representations such as point clouds or textured meshes, omnidirectional videos and light fields have also benefited from rendering tools first developed in the computer graphics field. Furthermore, when considering immersive video technologies, geometry and appearance are not computer generated, but rather estimated from images of a real world scene.

1.3.3 Imaging modalities

Considering what is discussed above for the immersiveness of video technologies, one can identify many imaging (which includes display as well) modalities. The following modalities can be considered among immersive video technologies, as they enrich the traditional video and provide a more realistic or immersive experience: high dynamic range (HDR) video, high frame rate (HFR) video, stereoscopic (or multiview) video, hologram technology, omnidirectional video, light fields, and volumetric video.

These modalities do increase the realism or sense of presence through their unique

ways. Nevertheless, referring to our discussion on realism and interactivity in Section 1.2.2, we can consider the interactivity vs. realism graph as shown in Figure 1.5. Some of the modalities mentioned above have high realism, but they might not have high interactivity, for example HDR and HFR technologies. On the other hand, the low-polygon computer graphics models in virtual reality applications are very interactive, but they lack realism. Taking only the modalities with high interactivity and high realism, we end up with five main imaging modalities: multiview video, holography, omnidirectional video, light fields, and volumetric video.

Since both multiview video and holography have been thoroughly studied in the past in many scientific articles and books, this book only focuses on three main imaging modalities: omnidirectional video, light fields, and volumetric video. Each of these imaging modalities are briefly discussed below and each have a part of several chapters in this book.

1.3.4 Omnidirectional imaging

Omnidirectional imaging systems [16,17] capture different viewing angles over time. Typically, the 360-degree camera is fixed in space, and it captures the world around the camera. Therefore, the 7D plenoptic function becomes 2D: $P(\theta, \phi)$ (3D for omnidirectional video). Ideally, all the rays coming from different parts from the sphere are captured. However, in practice, there are physical limitations such as parts that cannot be captured well (e.g., north and south poles of the sphere – as either of the poles are generally used for the handle that keeps the cameras together) or the

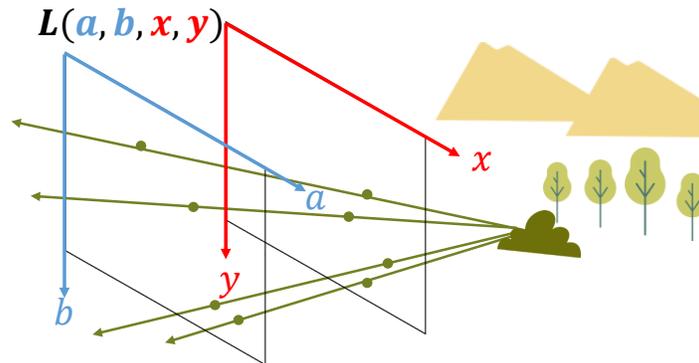


FIGURE 1.6

Light field imaging aims to consider the light as a field, much similar to electromagnetic fields. However, for practical reasons, it is mostly simplified. The most commonly used simplification is the two parallel planes parameterization of the 4D light field. In this context, the light rays are parameterized by their intersection with two parallel planes (a, b) and (x, y) .

need to stitch visual from various cameras placed.

1.3.5 Light field imaging

Light field has the broadest definition of all these imaging modalities as it aims to consider the light as a field, much similar to electromagnetic fields [18,19]. However, for practical reasons, it is mostly simplified. The most commonly used simplification is the two parallel planes parameterization of the 4D light field. In this context, the light rays are parameterized by their intersection with two parallel planes (a, b) and $(x, y)^2$, i.e. there is a one to one mapping between (a, b, x, y) and (x, y, θ, ϕ) , and the 7D plenoptic function can be reduced to the 4D light field $L(a, b, x, y)$ for static scenes as illustrated in 1.6. The time dimension t needs to be added when considering light field videos. Once captured, the original scene can be rendered for the user with realistic parallax, increasing the immersion, by giving the depth information to the viewer. Developments in the recent light field capturing systems [20] use spherical light field parametrization and show that the light field imaging systems can be utilized to capture and render panoramic scenes with 3DoF+. This illustrates that the different modalities presented in this paper are part of a continuum rather than isolated concepts.

1.3.6 Volumetric 3D imaging

Volumetric 3D imaging systems [21–23] retain most of the variables from the original 7D plenoptic function as the capture systems for this imaging modality needs capturing from various angles around the object in focus. For this purpose, dedicated studios are built and cameras are placed around the edges of these studios to capture what is placed in the center. Although these dedicated studios are required for high quality capture and content creation, it is also possible to make use of handheld cameras to capture volumetric content [23]. The static 3D contents and scenes can be captured casually using a single handheld camera, and multiple handheld cameras can be used for the capture of dynamic volumetric content outside dedicated studios.

A recent and fast growing trend is to use neural networks to represent a scene, called neural representation [24,25]. In such representation, the neural network typically serves as a continuous estimate of a function taking the position and orientation as input, and evaluating the corresponding scene color and transparency as output. Such representations enable high quality 6DoF rendering. By using a differentiable rendering process, the neural network representing the scene can be trained from a set of calibrated input views.

²Note that the notations in this book differ from tradition light field notations (s,t,u,v), see chapter ??

1.4 Challenges in the processing and content delivery pipeline

Similar to traditional imaging pipelines, the immersive imaging pipeline starts with the image acquisition of the real world objects. After the image acquisition, a processing step is generally required to transform the raw images into a desired representation. The content can then be compressed for storage or streaming. Unlike traditional 2D images, immersive video content is not meant to be visualized all at once by the user. This requires the design of specific rendering algorithms, streaming strategies, and / or dedicated visualization hardware. Finally, the quality of the novel immersive video content needs to be evaluated, knowing that this can be impacted by any of the previous steps.

1.4.1 Acquisition and representations

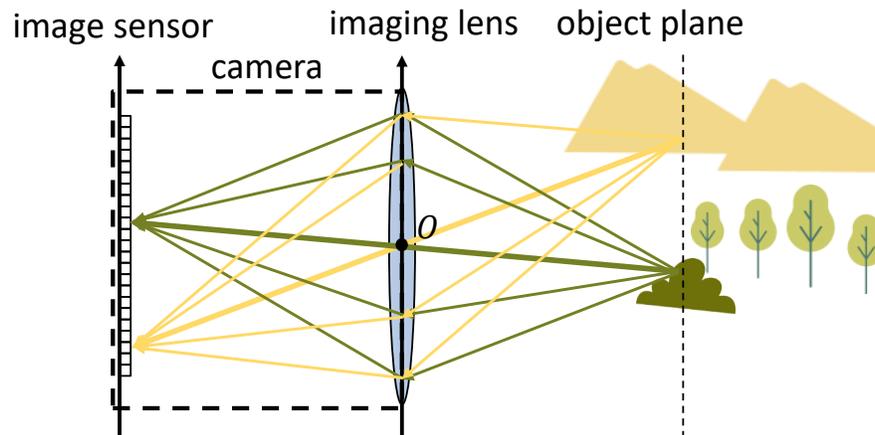
We saw in the previous section that the ability to provide an immersive imaging experience comes from the ability to sample the plenoptic function, in particular sampling light rays coming from multiple angles or directions. For this purpose, multiple acquisition devices and systems have been designed, which we review in this section.

1.4.1.1 Acquisition of 2D images and videos

As some of these systems rely on traditional 2D cameras, we first give here a reminder about traditional 2D cameras. As illustrated in Figure 1.7, a 2D camera can be modeled by a main lens and an image sensor. Light rays emitted from objects in multiple directions are thus re-converged through the lens on the image sensor, which is similar to having one single ray combining the intensity of all rays going through the lens optical center. For this reason, 2D camera can not directly capture immersive imaging content, as the angular information is lost. The pinhole camera model [26] is usually adopted for 2D cameras. Such model is essential to make a connection between pixels and world coordinates. Precise calibration procedures are required in order to obtain accurate model parameters.

1.4.1.2 Acquisition of immersive imaging data

There are two main categories of systems designed to capture immersive imaging content with regular 2D cameras: either use a single camera which is moved to capture different viewpoints of the scene or object of interests, or use multiple cameras rigged together and pointing in different directions. The first solution is usually simpler to implement, and allows for dense sampling of the plenoptic functions, as very close viewpoint images can be captured. However, it is by design limited to static scenes. To capture video contents, camera rigs have to be used, which are technically more challenging to implement and more expensive. It is also limited to a sparse sampling of the plenoptic function, as the camera casings mechanically prevent close viewpoints

**FIGURE 1.7**

2D lens camera model. Light rays arriving from different angles are all integrated on the same pixel of the image sensor, as if a single ray was going through the optical centre of the lens.

to be captured.

Existing systems adopt different camera geometries depending on the targeted immersive imaging modality: arranging outward looking cameras on a sphere can be used to capture omnidirectional videos or spherical light fields. Using inward looking cameras on a sphere can also be used to capture spherical light fields, and in theory volumetric videos. The most popular light field representation is based on the two parallel plane parameterization, which can be captured by arranging the cameras on a plane. Note that volumetric video usually does not impose any specific geometry on the camera positions, but does require the knowledge of the camera model parameters, i.e. for the pinhole camera model, intrinsics and extrinsics parameters.

Some specific camera designs beyond the traditional 2D cameras have also been proposed. Early works on curved mirror-based cameras or using fisheye lenses have been proposed to capture omnidirectional videos. More recently, lenslet arrays have been added to regular cameras to capture light fields. Note that specific models have to be derived for such cameras. Ongoing research is being carried out which combines multiple of the specific camera designs presented above in order to increase the immersion and provide full 6DoF content, e.g. use an array of omnidirectional camera [27], or multiple plenoptic cameras [28]. One of the challenge of such systems is their calibration in order to obtain an accurate estimate of their model parameters, which is more difficult than for the 2D pinhole camera model.

1.4.1.3 From raw data to content creation

Following acquisition, the captured raw images need to be processed in order to create the target immersive content and fit the target modality representation. Different processing methods are needed depending on the immersive imaging modality.

Several low level image processing steps are common to most modalities and extend the traditional 2D image processing techniques. This can include, debayering the raw image in order to obtain color RGB images, de-noising, or super-resolution. When creating immersive imaging content, additional care has to be taken to enforce the consistency of the processing among the different viewpoints. Furthermore, as mentioned in the previous section, a calibration step is often required to estimate the parameters of the model associated with the camera used for acquisition. Typically, parameters of the pinhole camera model for data captured with regular 2D cameras can be estimated using Structure-from-Motion (SfM) [26]. Calibration is also required for other processing methods such as depth estimation (see ??). Depth estimation is an essential step of volumetric video content creation, but can also provide additional data for further post-production and processing tasks, especially for light fields or stereo omnidirectional imaging. The estimated depth can be useful for various computer vision tasks such as segmentation, scene understanding, and view synthesis.

The different modalities discussed in this book may also require more specific processing. This typically includes stitching for omnidirectional imaging, rectifying the viewpoints images for light fields, and 3D reconstruction for volumetric imaging as mentioned above.

As part of content generation stage, recent learning-based approaches allow to create novel neural representations for enhanced rendering operations.

1.4.2 Compression and transmission

1.4.2.1 Compression of immersive videos

Due to their massive data size, cost-efficient coding solutions are required to store and deliver immersive videos. The majority of existing immersive video coding solutions are based on the video coding standards; high efficient video coding (HEVC) [29] and versatile video coding (VVC). Each coding standard aims to achieve approximately 50% bitrate reduction over its predecessor coding standard. For instance, VVC has achieved up to 50% bitrate reduction compared to HEVC by implementing new improvements for hybrid prediction/transform coding scheme and a set of new tools [30].

Omnidirectional video (ODV) is stored and transmitted with its 2D planar representations, containing redundant pixels. Several cost-efficient viewport-based coding techniques exist to exploit these redundant pixels, which VR devices do not use [31]. For instance, coding-friendly representations, such as cube map projections (CMP) and truncated square pyramid (TSP), achieve a cost-efficient coding performance.

Light fields consists in a collection of images, which exhibit a lot of self-similarities. These redundancies can be used to design cost-efficient compression

algorithms to help reduce the light field substantial volume of data. For instance, sparse subset coding techniques select only a subset of the light field views for encoding, and they reconstruct the remaining views at the decoder.

Volumetric video coding standard solutions is typically divided into two parts based on the used representations, namely, mesh-based and point-cloud based compression techniques. For point cloud compression, two different point cloud coding techniques are commonly used for volumetric video compression: geometry-based PCC (G-PCC) and video-based PCC (V-PCC) [32].

Immersive video standardization activities have recently been started with ISO/IEC MPEG Immersive Video (MIV) standard [7]. The upcoming MIV standardization aims to provide the capability to compress a given 3D scene representation captured by multiple real or virtual cameras. The MIV coding framework is designed to utilize multiview video plus depth representation of immersive video data to enable 3D reconstruction with 6DoF capability. A test model framework, which consists of reference software encoder, decoder, and renderer, was developed for immersive video during this standardization activity.

1.4.2.2 Streaming of immersive videos

State-of-the-art video communication technologies are adaptive bit-rate (ABR) streaming system, which is designed to deliver a given captured video to the users in the most cost-efficient way possible and with the highest quality. In the ABR streaming, a given captured video is prepared in several representations, encoded at different bit-rates, resolution, and frame-rate. Each video representation is divided into a number of equal duration segments. Different segments are aligned such that the client device can request an appropriate segment of video representation based on the network bandwidth. Figure 1.8 illustrates a schematic diagram for adaptive video delivery system.

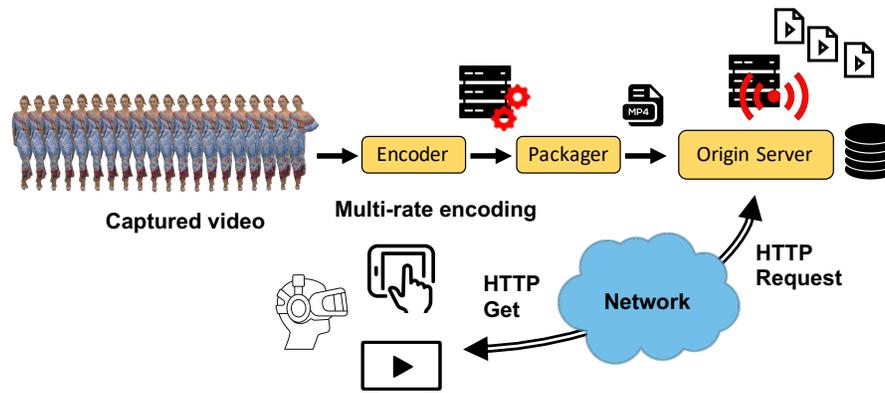


FIGURE 1.8

Schematic diagram of adaptive video delivery system.

In omnidirectional video streaming, only a small portion of a given content, called viewport, is used by the HMDs. Therefore, a very high resolution of the omnidirectional video is needed to deliver a perceptually acceptable quality level. To deploy omnidirectional video in adaptive streaming frameworks, the MPEG created the omnidirectional media format (OMAF) [33] which has two editions.

Due to its interactive use cases, cost-efficient streaming and transmission delay play important roles in light field streaming. To achieve cost-efficient transmission, most of the studies use approaches similar to sparse subset light field coding, wherein only a subset of the views can be transmitted to clients. In particular, transmission delay is an essential aspect that should be taken into account in light field streaming due to its interactive use cases, as studied in [34].

The volumetric video introduces additional challenges in streaming frameworks as real-time interaction with the content becomes crucial. Adaptive streaming strategies for point cloud representations have been investigated to optimize volumetric video streaming systems. For instance, in [35], various point cloud density levels were generated using spatial sub-sampling algorithms per given 3D point cloud content. The experimental results show that sub-sampling-based rate adaptation can significantly save point cloud streaming bandwidth.

1.4.2.3 Challenges for immersive video

To enable content delivery over the Internet (i.e., online) or any other medium (e.g., offline, stored media), compression and transmission techniques are needed to be developed. The presentation of the content itself at the receiver side requires specific rendering algorithms. Moreover, special display devices are used to present the visual output of these immersive imaging systems. Although there are lossless alternatives, most of the compression and transmission techniques introduce some form of perceptual or non-perceptual losses, and the rendering algorithm can also introduce visual artifacts. To ensure the highest quality of experience for the viewers, several quality assessment and visual attention mechanisms are used both during the compression and transmission and during the presentation of the media afterwards.

1.4.3 Rendering and display

As described above, digital imaging technologies rely on tightly packed and very structured grid of pixels. The pixel values that are already processed are passed to the rendering and display step, and converted into light at the display device. The pixel values are already given to the display driver, and there is little in terms of rendering for traditional video.

Since most of the rendering is already done prior to video distribution, rendering in traditional images and video usually refers to the techniques that determines the location within the screen, the spatial resolution, and the temporal resolution of the video. This includes solving problems such as subsampling and anti-aliasing.

Similarly, display is rather straightforward compared to other techniques that are

used in immersive imaging technologies. There are different display technologies including liquid crystal display (LCD), plasma display panel (PDP), and organic light-emitting diode (OLED) displays. Although technologies differ in terms of the presence of backlight, light leakage from different angles, and the contrast ratios, essentially, all display technologies for traditional video rely on using the pixel grid with minimal change in the location or spatial resolution of the pixels as mentioned above. Except for a small number of curved and/or foldable displays, all of them are essentially 2D planes.

Due to their unique requirements and different representation structures, immersive video technologies need specialized rendering methods to either render the spherical view onto the planar display, find the corresponding viewpoint for the light fields by image-based rendering, or render and rasterize the 3D models of the volumetric video. The techniques for renderings and novel display technologies are discussed in the following chapters, in each part for the three different display modalities.

1.4.4 Quality assessment

Images and video are versatile and can be used for many different tasks in addition to human consumption. Since immersive imaging technologies target recreation of the real world through digital imaging systems [36], end-users for immersive video technologies are human viewers. In this context, quality is described as “the outcome of an individual’s comparison and judgment process. It includes perception, reflection about the perception, and the description of the outcome” [37]. Moreover, the Quality of Experience (QoE) term is defined as “the degree of delight or annoyance of the user of an application or service” both in general sense [37] and for immersive media technologies [36].

Numerical quality scores can be obtained via psychophysical user studies (also known as subjective quality experiments) or quality estimation algorithms (also known as objective quality metrics). Briefly, the subjective quality assessment provide ground truth ratings of visual quality, while being time and cost expensive. The latter, objective quality metrics, are easy to execute, but they only offer computational predictions of visual quality.

1.4.4.1 Subjective quality assessment

As per its definition, quality is subjective. In subjective quality experiments, viewers (i.e., participants) are presented with visual stimuli and asked to provide their response to the experiment question. The collected data can be treated as ground truth in many cases; however, if not carefully planned, the tests may result in misleading and noisy data due to various psychological effects [38]. To minimize the noise in these tests, throughout years, there have been many efforts to create standards and recommendations by standardization committees (e.g., International Telecommunications Union – ITU) or groups formed by scientific experts (e.g., Moving Picture Experts Group – MPEG) [39–42].

The subjective test mainly fall into one if two categories: rating and ranking. Rating tests ask participants to “rate” (i.e., determine the quality of) the visual stimulus presented to them. Ranking tests ask participants to provide an order of preference. In rating tests, the participants can be presented with either a single stimulus, double stimuli, or multiple stimuli. The single stimulus (SS) methodologies generally ask to determine the inherent quality of the presented visual. Absolute Category Rating (ACR) [40], ACR with Hidden Reference (ACR-HR) [40], and Single Stimulus Continuous Quality Evaluation (SSCQE) [39,41] are three commonly used SS rating methodologies. The double stimulus (DS) methodologies present two visuals at the same time. Degradation category rating (DCR) [40], Double Stimulus Impairment Scale (DSIS) [41], Double-Stimulus Continuous Quality Evaluation (DSCQE) [39,41] are three commonly used DS methodologies. There can be also multiple stimulus presented to the viewer at the same time, of which Subjective Assessment Methodology for Video Quality (SAMVIQ) [42] is an example. In ranking tests, the most simple methodology is simply ranking multiple visuals, followed by statistical analysis. A very popular ranking method in signal processing and computer graphics communities is pairwise comparisons (PWC) methodology [40], in which the viewer is asked which option they prefer the more from two simultaneously presented stimuli. But, in some cases, a psychometric scaling may be necessary, which has its own challenges [43].

After data collection, to remove noisy data or to understand whether the results have meaningful difference, statistical analysis methods are used in most cases. These include outlier detection and removal [41], analysis of variance (ANOVA), pooling individual subjective quality scores, finding mean and standard deviation, and finding confidence intervals. These steps ensure that the collected data are cleaned and any bias or erroneous data are removed before their use.

1.4.4.2 Objective quality metrics

As mentioned above, subjective quality tests require meticulous effort in planning, execution, and analysis of results of the psychophysical user studies. Conducting such studies are not feasible for many applications, including compression, transmission, and other tasks that needs to be completed in real-time. For this purpose, objective quality metrics are used to estimate the visual quality of a given video.

Quality metrics work on the pixel values of video frames. Metrics can have different approaches for traditional video. Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) focus on the pixel error values. Structural Similarity Index Measure (SSIM [44]) and other structural similarity-based metrics (e.g., MS-SSIM [45], FSIM [46], etc.) leverage human visual system principles which mainly focus on identifying similarities between luminance and contrast of the two images. Information Fidelity Criterion (IFC) [47] and Visual Information Fidelity (VIF) [48] rely on the statistical characteristics of the image extracted using information theory principles. Some hybrid metrics such as Video Quality Metric (VQM) [49] and Video Multi-Assessment Fusion (VMAF) [50] combine various features to create

a single estimation measure for video. Other recent methods use neural network approaches [51].

The estimated quality values generally have numerical values that may be arbitrarily different from metric to metric. Therefore, when using these metrics in real applications or when bench-marking their performances, a fitting can be done to the subjective quality scores' range (e.g., from 0.8-1 to 0-100). This fitting can be done either using polynomial functions or nonlinear functions [52]. For performance comparisons, a set of statistical performance metrics (which were recommended by ITU [53]) are widely used in the literature. These metrics are: Pearson Correlation Coefficient (PCC), Spearman's Rank Ordered Correlation Coefficient (SROCC), Root Mean Squared Error (RMSE), and Outlier Ratio (OR). Recently, other methods to evaluate objective metrics were proposed [52,54,55], based on reformulating the task from correlation to classification.

1.4.4.3 Challenges and limitations for immersive video technologies

In immersive video technologies, although the foundational data structures rely on legacy imaging systems and pixel-based grid structures, the new representations and perceptual limitations dictate a different set of challenges and limitations. For example, the spherical nature of omnidirectional video, combined with the characteristics of human visual perception, makes ODV quality assessment very different from traditional video quality assessment.

1.5 Discussion and Perspectives

Immersive imaging systems aim at capturing and reproducing the plenoptic function. We presented, in this chapter, the theoretic background of the immersive video technologies, technologies that affect all modalities, and a brief overview of the content delivery pipeline, from image acquisition to display and quality assessment. This book, and therefore this chapter, focuses on three current prominent modalities which are omnidirectional imaging, light field imaging, and volumetric imaging, i.e. point cloud and textured meshes. One of the fundamental components of all these systems and modalities is to capture light from different angles. These systems thus capture a lot of visual data, which require specific acquisition systems and more efficient compression methods for storage and streaming purposes. Furthermore, the visual information captured is not intended to be visualized all at once, but rather needs dedicated rendering algorithms which can be displayed on traditional 2D screens or more advanced immersive display systems, such as head mounted displays or light field displays.

One of the key technologies enabling rapid improvement of immersive imaging systems is deep learning, either as a tool to improve processing of classical immersive imaging modalities, or as a novel immersive imaging representation.

As discussed in the following chapters, there has been significant advances in immersive imaging technologies in recent years. Nevertheless, several challenges still

lie ahead. These challenges include the popularization of immersive imaging systems, improving the reproduction fidelity of the captured real world content by developing better display and rendering systems, increasing the speed of 3D reconstruction for volumetric imaging systems to enable real-time 3D content creation, understanding and leveraging user behavior to increase the efficiency in adaptive streaming and compression, and more efficient compression and transmission systems to deal with the increasing amount of information captured by the immersive imaging systems.

Bibliography

- [1] M. Minsky, Telepresence, *Omni* (1980) 45–51.
- [2] W. A. IJsselstein, H. De Ridder, J. Freeman, S. E. Avons, Presence: concept, determinants, and measurement, in: *Human Vision and Electronic Imaging V*, Vol. 3959, International Society for Optics and Photonics, 2000, pp. 520–529.
- [3] J. Takatalo, G. Nyman, L. Laaksonen, Components of human experience in virtual environments, *Computers in Human Behavior* 24 (1) (2008) 1–15.
- [4] M. Lombard, T. Ditton, At the heart of it all: The concept of presence, *Journal of Computer-Mediated Communication* 3 (2) (1997).
- [5] P. Schelkens, T. Ebrahimi, A. Gilles, P. Gioia, K.-J. Oh, F. Pereira, C. Perra, A. M. G. Pinheiro, JPEG Pleno: Providing representation interoperability for holographic applications and devices, *ETRI Journal* 41 (1) (2019) 93–108. doi:10.4218/etrij.2018-0509.
- [6] M. Domański, O. Stankiewicz, K. Wegner, T. Grajek, Immersive visual media - MPEG-I: 360 video, virtual navigation and beyond, in: *International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2017.
- [7] J. M. Boyce, R. Doré, A. Dziembowski, J. Fleureau, J. Jung, B. Kroon, B. Salahieh, V. K. M. Vadakital, L. Yu, MPEG immersive video coding standard, *Proceedings of the IEEE* (2021) 1–16doi:10.1109/JPROC.2021.3062590.
- [8] F. Biocca, B. Delaney, Immersive virtual reality technology, in: F. Biocca, M. R. Levy (Eds.), *Communication in the Age of Virtual Reality*, Lawrence Erlbaum Associates, Inc., 1995, pp. 57—124).
- [9] H. G. Hoffman, J. Prothero, M. J. Wells, J. Groen, Virtual chess: Meaning enhances users' sense of presence in virtual environments, *International Journal of Human-Computer Interaction* 10 (3) (1998) 251–263.
- [10] U. Schultze, Embodiment and presence in virtual worlds: a review, *Journal of Information Technology* 25 (4) (2010) 434–449.
- [11] Cambridge English Dictionary, immerse, Online: <https://dictionary.cambridge.org/dictionary/english/immerse> (2022).
- [12] S. T. Coleridge, *Biographia literaria*, chapter xiv, West Sussex, England: Littlehampton Book Services 1975 (1817).
- [13] P. Milgram, H. Takemura, A. Utsumi, F. Kishino, Augmented reality: A class of displays on the reality-virtuality continuum, in: *Telemanipulator and telepresence technologies*, Vol. 2351, International Society for Optics and Photonics, 1995, pp. 282–292.
- [14] L.-H. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, A. Kumar, C. Bermejo, P. Hui, All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda, *arXiv preprint arXiv:2110.05352* (2021).
- [15] E. H. Adelson, J. R. Bergen, The plenoptic function and the elements of early vision, in: *Computational Models of Visual Processing*, MIT Press, 1991, pp. 3–20.

- [16] Y. Yagi, Omnidirectional sensing and its applications, *IEICE Transactions on Information and Systems* 82 (3) (1999) 568–579.
- [17] K. K. Sreedhar, I. D. D. Curcio, A. Hourunranta, M. Lepistö, Immersive media experience with MPEG OMAF multi-viewpoints and overlays, in: *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020, p. 333–336.
- [18] I. Ihrke, J. Restrepo, L. Mignard-Debise, Principles of light field imaging: Briefly revisiting 25 years of research, *IEEE Signal Processing Magazine* 33 (5) (2016) 59–69.
- [19] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, Y. Liu, Light field image processing: An overview, *IEEE Journal of Selected Topics in Signal Processing* 11 (7) (2017) 926–954.
- [20] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, P. Debevec, Immersive light field video with a layered mesh representation, *ACM Trans. Graph.* 39 (4) (Jul. 2020). doi:10.1145/3386569.3392485.
- [21] A. Smolic, 3D video and free viewpoint video—from capture to display, *Pattern recognition* 44 (9) (2011) 1958–1968.
- [22] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, S. Sullivan, High-quality streamable free-viewpoint video, *ACM Trans. Graphics* 34 (4) (Jul. 2015). doi:10.1145/2766945.
- [23] R. Pagés, K. Amlianitis, D. Monaghan, J. Ondřej, A. Smolić, Affordable content creation for free-viewpoint video and VR/AR applications, *Journal of Visual Communication and Image Representation* 53 (2018) 192–201. doi:10.1016/j.jvcir.2018.03.012.
- [24] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, Y. Sheikh, Neural volumes: Learning dynamic renderable volumes from images, *ACM Transactions on Graphics* 38 (4) (2019) 65:1–65:14.
- [25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng, NeRF: Representing scenes as neural radiance fields for view synthesis, in: *Proceedings of the European Conference on Computer Vision*, 2020.
- [26] R. Hartley, A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, 2003.
- [27] T. Maugey, L. Guillo, C. L. Cam, FTV360: a multiview 360° video dataset with calibration parameters, in: *Proceedings of the 10th ACM Multimedia Systems Conference*, 2019, pp. 291–295.
- [28] O. Johannsen, A. Sulc, B. Goldluecke, On linear structure from motion for light field cameras, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 720–728.
- [29] M. Wien, J. M. Boyce, T. Stockhammer, W.-H. Peng, Standardization status of immersive video coding, *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9 (1) (2019) 5–17.
- [30] N. Sidaty, W. Hamidouche, O. Déforges, P. Philippe, J. Fournier, Compression performance of the versatile video coding: Hd and uhd visual quality monitoring, in: *2019 Picture Coding Symposium (PCS)*, IEEE, 2019, pp. 1–5.
- [31] R. Shafi, W. Shuai, M. U. Younus, 360-Degree video streaming: A survey of the state of the art, *Symmetry* 12 (9) (2020) 1491.
- [32] D. B. Graziosi, O. Nakagami, S. Kuma, A. Zaghetto, T. Suzuki, A. Tabatabai, An overview of ongoing point cloud compression standardization activities: video-based (V-PCC) and geometry-based (G-PCC), *APSIPA Transactions on Signal and Information Processing* 9 (2020).
- [33] ISO/IEC, Information technology — Coded representation of immersive media — Part 2: Omnidirectional media format, Standard, International Organization for Standardization (Jan. 2019).
- [34] M. Alain, C. Ozcinar, A. Smolic, A study of light field streaming for an interactive refocusing application, in: *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, pp. 3761–3765.
- [35] M. Hosseini, C. Timmerer, Dynamic adaptive point cloud streaming, in: *Proceedings of the 23rd Packet Video Workshop*, 2018, pp. 25–30.

- [36] A. Perkis, C. Timmerer, S. Baraković, J. Baraković Husić, S. Bech, S. Bosse, J. Botev, K. Brunnström, L. Cruz, K. De Moor, A. de Polo Saibanti, W. Durnez, S. Egger-Lampl, U. Engelke, T. H. Falk, J. Gutiérrez, A. Hameed, A. Hines, T. Kojic, D. Kukulj, E. Liotou, D. Milovanovic, S. Möller, N. Murray, B. Naderi, M. Pereira, S. Perry, A. Pinheiro, A. Pinilla, A. Raake, S. R. Agrawal, U. Reiter, R. Rodrigues, R. Schatz, P. Schelkens, S. Schmidt, S. S. Sabet, A. Singla, L. Skorin-Kapov, M. Suznjevic, S. Uhrig, S. Vlahović, J.-N. Voigt-Antons, S. Zadtootaghaj, QUALINET white paper, on definitions of immersive media experience (IMEx), European Network on Quality of Experience in Multimedia, Systems and Services, 14th QUALINET meeting (online), online: <https://arxiv.org/abs/2007.07032> (May 2020).
- [37] P. Le Callet, S. Möller, A. Perkis, Qualinet white paper on definitions of quality of experience, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2 (Mar 2013).
- [38] F. De Simone, Selected contributions on multimedia quality evaluation, Ph.D. thesis, École Polytechnique Fédérale de Lausanne (2012).
- [39] T. Alpert, J. Evain, Subjective quality evaluation: the SSCQE and DSCQE methodologies, EBU Technical Review (1997).
- [40] ITU-T, Subjective video quality assessment methods for multimedia applications, ITU-T Recommendation P.910 (Apr 2008).
- [41] ITU-R, Methodology for the subjective assessment of the quality of television pictures, ITU-R Recommendation BT.500-13 (Jan 2012).
- [42] F. Kozamernik, V. Steinmann, P. Sunna, E. Wyckens, SAMVIQ-A new EBU methodology for video quality evaluations in multimedia, SMPTE Motion Imaging Journal 114 (4) (2005) 152–160.
- [43] E. Zerman, V. Hulusic, G. Valenzise, R. K. Mantiuk, F. Dufaux, The relation between MOS and pairwise comparisons and the importance of cross-content comparisons, in: IS&T Electronic Imaging, Human Vision and Electronic Imaging XXII, 2018.
- [44] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.
- [45] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: 37th Asilomar Conference on Signals, Systems Computers, Vol. 2, IEEE, 2003, pp. 1398–1402.
- [46] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Transactions on Image Processing 20 (8) (2011) 2378–2386. doi:10.1109/TIP.2011.2109730.
- [47] H. R. Sheikh, A. C. Bovik, G. De Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Transactions on Image Processing 14 (12) (2005) 2117–2128.
- [48] H. R. Sheikh, A. C. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2) (2006) 430–444.
- [49] M. H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Transactions on Broadcasting 50 (3) (2004) 312–322. doi:10.1109/TBC.2004.834028.
- [50] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara, Toward a practical perceptual video quality metric, <https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652> (Jan 2019).
- [51] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, Z. Bai, C3DVQA: Full-reference video quality assessment with 3D convolutional neural network, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 4447–4451.
- [52] ITU-T, Method for specifying accuracy and cross-calibration of video quality metrics (VQM), ITU-T Recommendation J.149 (Mar 2004).
- [53] ITU-T, Methods, metrics and procedures for statistical evaluation, qualification and comparison of

- objective quality prediction models, ITU-T Recommendation P.1401 (Jul 2012).
- [54] L. Krasula, K. Fliegel, P. Le Callet, M. Klíma, On the accuracy of objective image and video quality models: New methodology for performance evaluation, in: 8th International Conference on Quality of Multimedia Experience (QoMEX), IEEE, 2016.
 - [55] E. Zerman, G. Valenzise, F. Dufaux, An extensive performance evaluation of full-reference HDR image quality metrics, *Quality and User Experience* 2 (5) (2017). doi:10.1007/s41233-017-0007-4. URL <http://dx.doi.org/10.1007/s41233-017-0007-4>