1

Introduction

1.1 Why 3D Communications?

Thanks to the great advancement of hardware, software, and algorithms in the past decade, our daily life has become a major digital content producer. Nowadays, people can easily share their own pieces of artwork on the network with each other. Furthermore, with the latest development in 3D capturing, signal processing technologies, and display devices, as well as the emergence of 4G wireless networks with very high bandwidth, coverage, and capacity, and many advanced features such as quality of service (QoS), low latency, and high mobility, 3D communication has become an extremely popular topic. It seems that the current trend is closely aligned with the expected roadmap for reality video over wireless, estimated by Japanese wireless industry peers in 2005 (as shown in Figure 1.1), according to which the expected deployment timing of stereo/multi-view/hologram video is around the same time as the 4G wireless networks deployment. Among those 3D video representation formats, the stereoscopic and multi-view 3D videos are more mature and the coding approaches have been standardized in Moving Picture Experts Group (MPEG) as "video-plus-depth" (V+D) and the Joint Video Team (JVT) Multi-view Video Coding (MVC) standard, respectively. The coding efficiency study shows that coded V+D video only takes about 1.2 times bit rate compared to the monoscopic video (i.e., the traditional 2D video). Clearly, the higher reality requirements would require larger volumes of data to be delivered over the network, and more services and usage scenarios to challenge the wireless network infrastructures and protocols.

From a 3D point of view, reconstructing a scene remotely and/or reproducibly as being presented face-to-face has always been a dream through human history. The desire for such technologies has been pictured in many movies, such as *Star Trek*'s Holodeck, *Star War*s' Jedi council meeting, *The Matrix*'s matrix, and *Avatar*'s Pandora. The key technologies to enable such a system involve many complex components, such as a capture system to describe and record the scene, a content distribution system to store/transmit the recorded scene, and a scene reproduction system to show the captured scenes to end users. Over the past several decades, we have witnessed the success of many applications, such as television broadcasting systems in analog (e.g., NTSC, PAL) and digital (e.g., ATSC, DVB) format, and home entertainment system in VHS, DVD, and Blu-ray format.

3D Visual Communications, First Edition. Guan-Ming Su, Yu-Chi Lai, Andres Kwasinski and Haohong Wang.

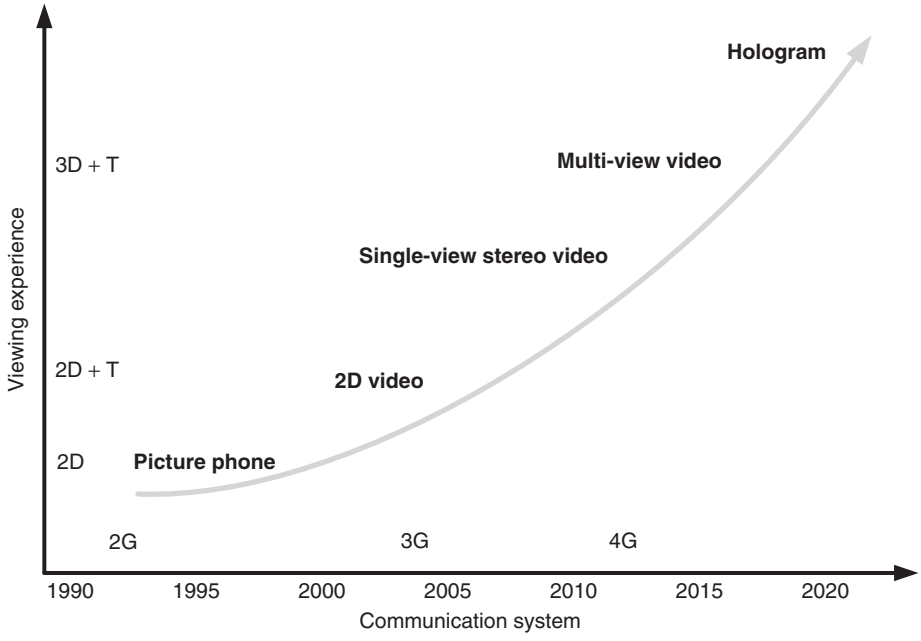

Figure 1.1 Estimated reality video over wireless development roadmap.

Although those systems have served for many years and advanced in many respects to give better viewing experiences, end users still feel that the scene reconstruction has its major limitation: the scene presentation is on a 2D plane, which significantly differs from the familiar three-dimensional view of our daily life. In a real 3D world, humans can observe objects and scenes from different angles to acquire a better understanding of the geometry of the watched scenes, and nonverbal signals and cues in visual conversation. Besides, humans can perceive the depth of different objects in a 3D environment so as to recognize the physical layout and location for each object. Furthermore, 3D visual systems can provide immersive viewing experience and higher interaction. Unfortunately, the existing traditional 2D visual systems cannot provide those enriched viewing experiences.

The earliest attempt to construct a 3D image was via the anaglyph stereo approach which was demonstrated by W. Rollmann in 1853 and J. C. D'Almeida in 1858 and patented in 1891 by Louis Ducos du Hauron. In 1922, the earliest confirmed 3D film was premiered at the Ambassador Hotel Theater in Los Angeles and was also projected in the red/green anaglyph format. In 1936, Edwin H. Land invented the polarizing sheet and demonstrated 3D photography using polarizing sheet at the Waldorf-Astoria Hotel. The first 3D golden era was between 1952 and 1955, owing to the introduction of color stereoscopy. Several golden eras have been seen since then. However, there are many factors affecting the popularity and success of 3D visual systems, including the 3D visual and content distribution technologies, the viewing experience, the end-to-end ecosystem, and competition from improved 2D systems. Recently, 3D scene reconstruction algorithms have achieved great improvement, which enables us to reconstruct a 3D scene from a 2D one and from stereoscope images, and the corresponding hardware can support the heavy computation

at a reasonable cost, and the underlying communication systems have advanced to provide sufficient bandwidth to distribute the 3D content. Therefore, 3D visual communication systems have again drawn considerable attention from both academia and industry.

In this book, we discuss the details of the major technologies involved in the entire end-to-end 3D video ecosystem. More specifically, we address the following important topics and the corresponding opportunities:

- the lifecycle of the 3D video content through the end-to-end 3D video communication framework,
- the 3D content creation process to construct a 3D visual experience,
- the different representations and compression formats for 3D scenes/data for content distribution. Each format has its own advantages and disadvantages. System designers can choose the appropriate solution for given the system resources, such as computation complexity and communication system capacity. Also, understanding the unequal importance of different syntaxes, decoding dependencies, and content redundancies in 3D visual data representation and coding can help system designers to adopt corresponding error resilient methods, error concealment approaches, suitable unequal error protection, and customized dynamic resource allocation to improve the system performance,
- the advanced communication systems, such as 4G networks, to support transmission of 3D visual content. Being familiar with those network features can help the system designer to design schedulers and resource allocation schemes for 3D visual data transmission over 4G networks. Also, we can efficiently utilize the QoS mechanisms supported in 4G networks for 3D visual communications,
- the effective 3D visual data transmission and network architectures to deliver 3D video services and their related innovative features,
- the 3D visual experience for typical users, the factors that impact on the user experiences, and 3D quality of experience (QoE) metrics from source, network, and receiver points of view. Understanding the factors affecting 3D QoE is very important and it helps the system designer to design a QoE optimized 3D visual communications system to satisfy 3D visual immersive expectations,
- the opportunities of advanced 3D visual communication applications and services, for example, how to design the source/relay/receiver side of an end-to-end 3D visual communication system to take advantage of new concepts of computing, such as green computing, cloud computing, and distributed/collaborated computing, and how to apply scalability concepts to handle 3D visual communications given the heterogeneous 3D terminals in the networks is an important topic.

1.2 End-to-End 3D Visual Ecosystem

As shown by the past experience and lessons learned from the development and innovation of visual systems, the key driving force is all about how to enrich the user experiences, or so-called QoE. The 3D visual system also faces the same issues. Although a 3D visual system provides a dramatic new user experience after traditional 2D systems, the QoE concept has to be considered at every stage of the communication system pipeline during system design and optimization work to ensure the worthwhileness of moving from 2D to

3D. There are many factors affecting the QoE, such as errors in multidimensional signal processing, lack of information, packet loss, and optical errors in display. Improperly addressing QoE issues will result in visual artifacts (objectively and subjectively), visual discomfort, fatigue, and other things that degrade the intended 3D viewing experiences.

An end-to-end 3D visual communication pipeline consists of the content creation, 3D representation, data compression, transmission, decompression, post-processing, and 3D display stages, which also reflects the lifecycle of a 3D video content in the system. We illustrate the whole pipeline and the corresponding major issues in Figure 1.2. In addition, we also show the possible feedback information from later stages to earlier stages for possible improvement of 3D scene reconstruction.

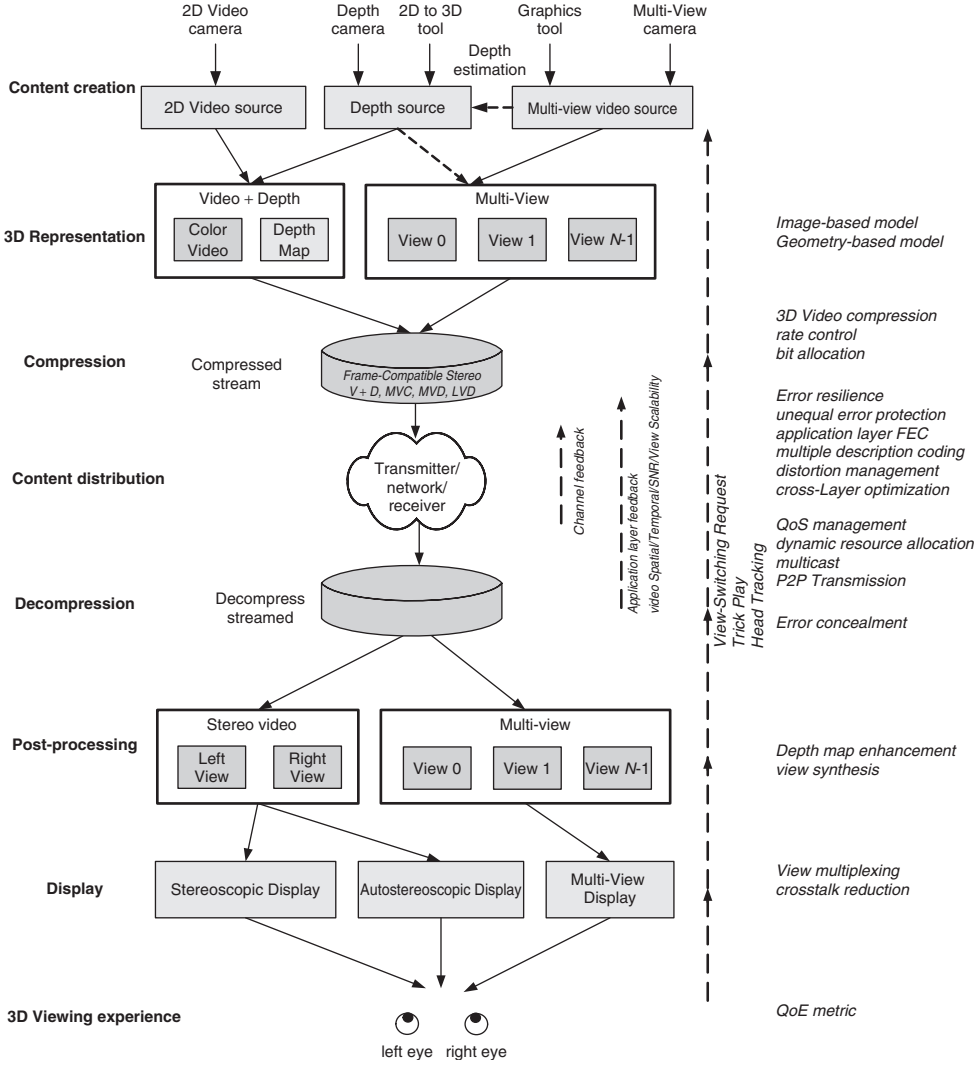

Figure 1.2 End-to-end 3D visual ecosystem.

The first stage of the whole pipeline is the content creation. The goal of the content creation stage is to produce 3D content based on various data sources or data generation devices. There are three typical ways of data acquisition which result in different types of data formats. The first is to use a traditional 2D video camera, which captures 2D images; the image can be derived for 3D data representation in the later stage of the pipeline. The second type is to use a depth video camera to measure the depth of each pixel corresponding to its counterpart color image. The registration of depth and 2D color image may be needed if sensors are not aligned. Note that in some depth cameras, the spatial resolution is lower than that of a 2D color camera. The depth image can also be derived from a 2D image with 2D-to-3D conversion tools; often the obtained depth does not have a satisfactory precision and thus causes QoE issues. The third type is to use an N-view video camera, which consists of an array of 2D video cameras located at different positions around one scene and all cameras are synchronized to capture video simultaneously, to generate N-view video. Using graphical tools to model and create 3D scene is another approach which could be time consuming, but it is popular nowadays to combine both graphical and video capturing and processing methods in the 3D content creation.

In the next stage, the collected video/depth data will be processed and transformed into 3D representation formats for different targeted applications. For example, the depth image source can be used in image plus depth rendering or processed for N-view application. Since the amount of acquired/processed/transformed 3D scene is rather large compared to single-view video data, there is a strong need to compress the 3D scene data. On the other hand, applying traditional 2D video coding schemes separately to each view or each different data type is inefficient as there exist certain representation/coding redundancies among neighboring views and different data types. Therefore, a dedicated compression format is needed at the compression stage to achieve better coding efficiency. In the content distribution stage, the packet loss during data delivery plays an important role in the final QoE, especially for streaming services. Although certain error concealment algorithms adopted in the existing 2D decoding and post-processing stages may alleviate this problem, directly applying the solution developed for 2D video system may not be sufficient. This is because the 3D video coding introduces more coding dependencies, and thus error concealment is much more complex compared to that in 2D systems. Besides, the inter-view alignment requirement in 3D video systems also adds plenty of difficulties which do not exist in 2D scenarios. The occlusion issue is often handled at the post-processing stage, and the packet loss will make the occlusion post-processing even more difficult. There are also some other application layer approaches to relieve the negative impact of packet loss, such as resilient coding and unequal error protection (UEP), and those technologies can be incorporated into the design of the 3D visual communication system to enrich the final QoE. At the final stage of this 3D visual ecosystem, the decoded and processed 3D visual data will be displayed on its targeted 3D display. Depending on the type of 3D display, each display has its unique characteristics of artifacts and encounters different QoE issues.

1.2.1 3D Modeling and Representation

3D scene modeling and representation is the bridging technology between the content creation, transmission, and display stages of a 3D visual system. The 3D scene modeling

and representation approaches can be classified into three main categories: geometry-based modeling, image based modeling, and hybrid modeling. Geometry-based representation typically uses polygon meshes (called surface-based modeling), 2D/3D points (called point-based modeling), or voxels (called volume-based modeling) to construct a 3D scene. The main advantage is that, once geometry information is available, the 3D scene can be rendered from any viewpoint and view direction without any limitation, which meets the requirement for a free-viewpoint 3D video system. The main disadvantage is in the computational cost of rendering and storing, which depends on the scene complexity, that is the total number of triangles used to describe the 3D world. In addition, geometry-based representation is generally an approximation to the 3D world. Although there are offline photorealistic rendering algorithms to generate views matching our perception of the real world, the existing algorithms using graphics pipeline still cannot produce realistic views on the fly.

The image based modeling goes to the other extreme, not using any 3D geometry, but using a set of images captured by a number of cameras with predesigned positions and settings. This approach tends to generate high quality virtual view synthesis without the effort of 3D scene reconstruction. The computation complexity via image based representation is proportional to the number of pixels in the reference and output images, but in general not to the geometric complexity such as triangle counts. However, the synthesis ability of image based representation has limitations on the range of view change and the quality depends on the scene depth variation, the resolution of each view, and the number of views. The challenge for this approach is that a tremendous amount of image data needs to be stored, transferred, and processed in order to achieve a good quality synthesized view, otherwise interpolation and occlusion artifacts will appear in the synthesized image due to lack of source data.

The hybrid approach can leverage these two representation methods to find a compromise between the two extremes according to given constraints. By adding geometric information into image based representation, the disocclusion and resolution problem can be relieved. Similarly, adding image information captured from the real world into geometry-based representation can reduce the rendering cost and storage. As an example, using multiple images and corresponding depth maps to represent 3D scene is a popular method (called depth image based representation), in which the depth maps are the geometric modeling component, but this hybrid representation can reduce the storage and processing of many extra images to achieve the same high-quality synthesized view as the image based approach. All these methods are demonstrated in detail in Chapters 2 and 4.

1.2.2 3D Content Creation

Other than graphical modeling approaches, the 3D content can be captured by various processes with different types of cameras. The stereo camera or depth camera simultaneously captures video and associated per-pixel depth or disparity information; the multi-view camera captures multiple images simultaneously from various angles, then multi-view matching (or correspondence) process is required to generate the disparity map for each pair of cameras, and then the 3D structure can be estimated from these disparity maps. The most challenging scenario is to capture 3D content from a normal 2D (or monoscopic) camera, which lacks of disparity or depth information, and where a

2D-to-3D conversion algorithm has to be triggered to generate an estimated depth map and thus the left and right views. The depth map can be derived from various types of depth cues, such as the linear perspective property of a 3D scene, the relationship between object surface structure and the rendered image brightness according to specific shading models, occlusion of objections, and so on. For complicated scenes, the interactive 2D-to-3D conversion, or offline conversion, tends to be adopted, that is, human interaction is required at certain stages of the processing flow, which could be in object segmentation, object selection, object shape or depth adjustment, object occlusion order specification, and so on. In Chapter 4, a few 2D-to-3D conversation systems are showcased to give details of the whole process flow.

1.2.3 3D Video Compression

Owing to the huge amount of 3D video data, there is a strong need to develop efficient 3D video compression methods. The 3D video compression technology has been developed for more than a decade and there have been many formats proposed. Most 3D video compression formats are built on state-of-the-art video codecs, such as H.264. The compression technology is often a tradeoff between the acceptable level of computation complexity and affordable budget in the communication bandwidth. In order to reuse the existing broadcast infrastructure originally designed for 2D video coding and transmission, almost all current 3D broadcasting solutions are based on a frame-compatible format via spatial subsampling approach, that is, the original left and right views are subsampled into half resolution and then embedded into a single video frame for compression and transmission over the infrastructure as with 2D video, and at the decoder side the demultiplexing and interpolation are conducted to reconstruct the dual views. The subsampling and merging can be done by either (a) side-by-side format, proposed by Sensio, RealD, and adopted by Samsung, Panasonic, Sony, Toshiba, JVC, and DirectTV (b) over/under format, proposed by Comcast, or (c) checkerboard format. A mixed-resolution approach is proposed, which is based on the binocular suppression theory showing that the same subjective perception quality can be achieved when one view has a reduced resolution. The mixed-resolution method first subsamples each view to a different resolution and then compresses each view independently.

Undoubtedly, the frame-compatible format is very simple to implement without changing the existing video codec system and underlying communication infrastructure. However, the correlation between left and right views has not been fully exploited, and the approach is mainly oriented to the two-view scenario but not to the multi-view 3D scenario. During the past decade, researchers have also investigated 3D compression from the coding perspective and 3D video can be represented in the following formats: two-view stereo video, video-plus-depth (V+D), multi-view video coding (MVC), multi-view video-plus-depth (MVD), and layered depth video (LDV). The depth map is often encoded via existing a 2D color video codec, which is designed to optimize the coding efficiency of the natural images. It is noted that depth map shows different characteristics from natural color image. Researchers have proposed several methods to improve the depth-based 3D video compression. In nowadays, free-viewpoint 3D attracts a lot of attention, in which the system allows end users to change the view position and angle to enrich their immersive experience. Hybrid approaches combining

geometry-based and image based representation are typically used to render the 3D scene for free-viewpoint TV. In Chapter 5, we discuss V+D, MVC, MVD, and LDV.

1.2.4 3D Content Delivery

Transmitting compressed 3D video bit streams over networks have more challenges than with conventional 2D video. From the video compression system point of view, the state-of-the-art 3D video codec introduces more decoding dependency to reduce the required bit rate due to the exploitation of the inter-view and synthesis prediction. Therefore, the existing mono-view video transmission scheme cannot be applied directly to these advanced 3D formats. From the communication system perspective, the 3D video bit stream needs more bandwidth to carry more views than the mono-view video. The evolution of cellular communications into 4G wireless network results in significant improvements of bandwidth and reliability. The end mobile user can benefit from the improved network infrastructure and error control to enjoy a 3D video experience. On the other hand, the wireless transmission often suffers frequent packet/bit errors; the highly bit-by-bit decoding-dependent 3D video stream is vulnerable to those errors. It is important to incorporate error correction and concealment techniques, as well as the design of an error resilient source coding algorithm to increase the robustness of transmitting 3D video bit streams over wireless environments. Since most 3D video formats are built up on existing 2D video codec, many techniques developed for 2D video systems can be extended or adapted to consider properties of 3D video. One technique that offers numerous opportunities for this approach is unequal error protection. Depending on the relative importance of the bit stream segments, different portions of the 3DTV bit stream are protected with different strengths of forward error control (FEC) codes. Taking the stereoscopic video streaming as an example, a UEP scheme can be used to divide the stream into three layers of different importance: intra-coded left-view frames (the most important ones), left-view predictive coded frames, and right-view frames encoded from both intra-coded and predictive left-view frames (the least valuable ones). For error concealment techniques, we can also draw from properties inherent to 3D video. Taking video plus depth format as an example, we can utilize the correlation between video and depth information to do error concealment. The multiple-description coding (MDC) is also a promising technology for 3D video transmission. The MDC framework will encode the video in several independent descriptions. When only one description is received, it can be decoded to obtain a lower-quality representation. When more than one description is received, they can be combined to obtain a representation of the source with better quality. The final quality depends on the number of descriptions successfully received. A simple way to apply multiple-description coding technology on 3D stereoscopic video is to associate one description with the right view and one with the left view. Another way of implementing multiple-description coding for 3D stereoscopic video and multi-view video consists of independently encoding one view and encoding the second view predicted with respect to the independently encoded view. This later approach can also be considered as a two-layer, base plus enhancement, encoding. This methodology can also be applied to V+D and MVD, where the enhancement layer is the depth information. Different strategies for advanced 3D video delivery over different content delivery path will be discussed in Chapters 8 and 10. Several 3D applications will be dealt with in Chapter 9.

1.2.5 3D Display

To perceive a 3D scene by the human visual system (HVS), the display system is designed to present sufficient depth information for each object such that HVS can reconstruct each object's 3D positions. The HVS recognizes objects' depth from the real 3D world through the depth cues. Therefore, the success of a 3D display depends on how well the depth cues are provided, such that HVS can observe a 3D scene. In general, depending on how many viewpoints are provided, the depth cues can be classified into monocular, binocular, and multi-ocular categories. The current 3DTV systems that consumers can buy in retail stores are all based on stereoscopic 3D technology with binocular depth cues. This stereoscopic display will multiplex two views at the display side and the viewers need to wear special glasses to de-multiplex the signal to get the left and right view. Several multiplexing/de-multiplexing approaches have been proposed and implemented in 3D displays, including wavelength division (color) multiplexing, polarization multiplexing, and time multiplexing.

For 3D systems without aided glasses, called auto-stereoscopic display (AS-D), the display system uses optical elements such as parallax barriers (occlusion-based approach) or lenticular lenses (refraction-based approach) to guide the two-view images to the left and right eyes of the viewer in order to generate the realistic 3D sense. In other words, the multiplexing and de-multiplexing process is removed compared to the stereoscopic display. Mobile 3DTV is an example of an AS-D product that we have seen in the market. The N-view AS-D 3DTVs or PC/laptop monitors have been in demos for many years by Philips, Sharp, Samsung, LG, Alioscopy, and so on, in which it explores the stereopsis of 3D space for multiple viewers without the need of glasses. However, the visual quality of these solutions still has lots of room to improve. To fully enrich the immersive visual experience, end users would want to interactively control the viewpoint, which is called free-viewpoint 3DTV (FVT). In a typical FVT system, the viewer's head and gaze are tracked to generate the viewing position and directions and thus to calculate images directed to the viewer's eyes. To render free-viewpoint video, the 3D scene needs to be synthesized and rendered from the source data in order to support the seamless view generation during the viewpoint changing.

To achieve full visual reality, holographic 3D display is a type of device to reconstruct the optical wave field such that the reconstructed 3D light beam can be seen as the physical presentation of the original object. The difference between conventional photography and holography is that photography can only record amplitude information for an object but holography attempts to record both the amplitude and phase information. Knowing that current image recoding systems can only record the amplitude information, holography needs a way to transform the phase information such that it can be recorded in an amplitude-based recoding system. For more details on 3D displays and their theory behind them, readers can refer to Chapter 3.

1.2.6 3D QoE

Although 3D video brings a brand new viewing experience, it does not necessarily increase the perceived quality if the 3D system is not carefully designed and evaluated. The 3D quality of experience refers to how humans perceive the 3D visual information, including

the traditional 2D color/texture information and the additional perception of depth and visual comfort factors. As the evaluation criteria to measure the QoE of 3D systems is still in its early stages, QoE-optimized 3D visual communications systems still remain an open research area. At the current stage, the efforts to address 3D QoE are considering the fidelity and comfort aspects. 3D fidelity evaluates the unique 3D artifacts generated and propagated through the whole 3D visual processing pipeline, and comfort refers to the visual fatigue and discomfort to the viewers induced by the perceived 3D scene.

In general, stereoscopic artifacts can be categorized as structure, color, motion, and binocular. Structure artifacts characterize those that affect human perception on image structures such as boundaries and textures, and include tiling/blocking artifacts, aliasing, staircase effect, ringing, blurring, false edge, mosaic patterns, jitter, flickering, and geometric distortion. The color category represents artifacts that affect the color accuracy, with examples including mosquito noise, smearing, chromatic aberration, cross-color artifacts, color bleeding, and rainbow artifacts. The motion category includes artifacts that affect the motion vision, such as motion blur and motion judder. Binocular artifacts represent those that affect the stereoscopic perception of the 3D world, for example, keystone distortion, cardboard effect, depth plane curvature, shear distortion, puppet theater effect, ghosting, perspective rivalry, crosstalk, depth bleeding, and depth ringing. Note that AS-D suffers more crosstalk artifacts than stereoscopic scenarios. This is mainly caused by imperfect separation of the left and right view images and is perceived as ghosting artifacts. The magnitude of crosstalk is affected by two factors: observing position between display and the observer and the quality of the optical filter in the display. The extreme case of crosstalk is the pseudoscopic (reversed stereo) image where the left eye sees the image representing the right view and the right eye sees the image representing the left view.

Free-viewpoint 3D systems have more distinctive artifacts due to the need of synthesizing new views from 3D scene representations. In a highly constrained environment camera parameters can be calibrated precisely and, as a result, visual artifacts in view synthesis arise principally from an inexact geometric representation of the scene. In an unconstrained environment where the lighting conditions and background are not fixed and the videos may have different resolution and levels of motion blur, the ambiguity in the input data and inaccuracies in calibration and matting cause significant deviation in a reconstructed view from the true view of the scene.

Visual fatigue refers to a decrease in the performance of the human vision system, which can be objectively measured; however, its subjective counterpart, visual discomfort, is hard to quantify. These factors affect whether end users enjoy the entire 3D experience and are willing to purchase 3D consumer electronic devices. In Chapter 7, more details on QoE topics will be discussed.

1.3 3D Visual Communications

Living in an era of widespread mobility and networking, where almost all consumer electronic devices are endpoints of the wireless/wired networks, the deployment of 3D visual representation will significantly challenge the network bandwidth as well as the computational capability of terminal points. In other words, the data volume received in an endpoint required to generate 3D views will be many times greater than that in a single view of a 2D system, and hence the new view generation process sets a higher requirement for the endpoint's computational capability. The emerging 4G networks can

significantly improve the bandwidth as well as include many new features designed specifically for high-volume data communications, which fit well into the timing of 3D visual communications. The two main 4G standards are 3GPP LTE (long-term evolution) [1] and IEEE 802.16m WiMAX (sometimes called WiMAX 2) [2]. Both standards provide features in the physical layer achieving high data rates and low latency based on orthogonal frequency-division multiplexing (OFDM) technology with adaptive modulation and coding (AMC) and multiple transmit/receive antenna (MIMO) support. In the upper communication layers, 4G networks support all-IP architectures and allow the uplink scheduler at the base station to learn the buffer status for the associated mobile devices by offering several quality of service (QoS) mechanisms to establish connections with different scheduling types and priorities.

In general, as the technologies, infrastructures and terminals evolving in wireless system (as shown in Figure 1.3) from 1G, 2G, 3G to 4G and from wireless LAN to broadband wireless access to 4G, the 4G system would contain all the standards that the earlier generations had implemented. Among the few technologies that are currently considered for 4G are 3GPP LTE, and WiMAX. Among the few technologies that are currently considered for 4G are 3GPP LTE and WiMAX. The details about the 4G networks will be presented in Chapter 6. We will discuss 3D over LTE and WiMAX in Chapter 10.

1.4 Challenges and Opportunities

As 3D visual communication has become one of the main focus research areas for the coming decade, we emphasize some potential research directions in this chapter. Streaming 3D video over 4G wireless networks has become a feasible and practicable application.

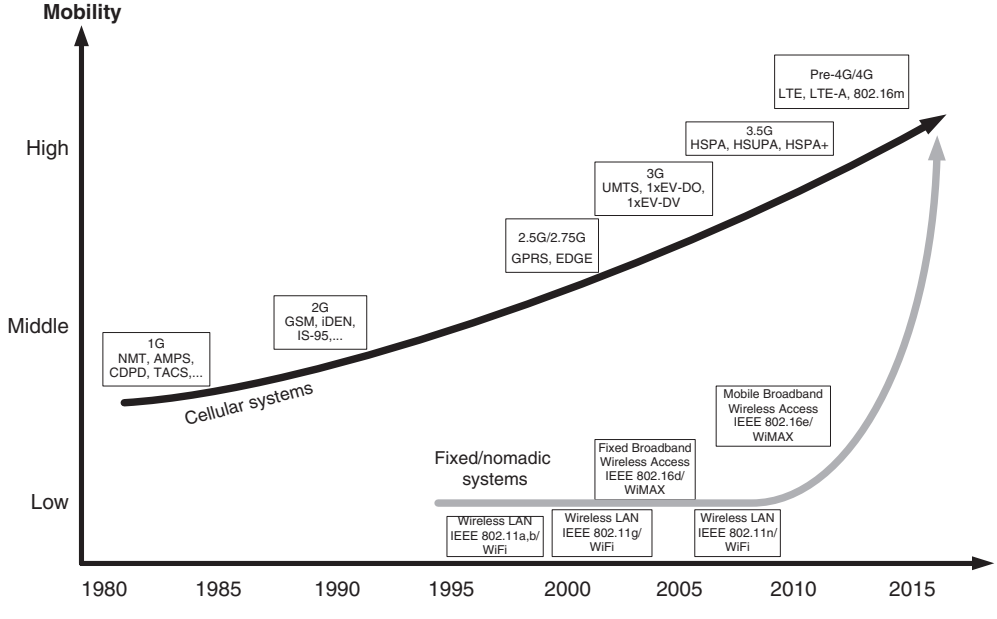

Figure 1.3 Evolving of wireless communication system.

4G wireless network standards include mechanisms to establish connections with different scheduling types and priorities, which support, for example, traffic with guaranteed maximum and minimum traffic rates. Also, the standards include different mechanisms that allow the uplink scheduler at the base station to learn the status of the buffers at the mobiles. Since the resource allocation mechanism is not yet specified in 4G wireless communication standard, it is an important topic to jointly consider the design of schedulers and resource allocation mechanisms with the QoS-related mechanisms in the standard to meet the final users' QoE requirement. The combination of the flexibility offered by 4G physical layer technologies, such as OFDM and AMC, and 4G features at higher layers, such as QoS support for heterogeneous services and all-IP, results in a framework that is suitable for delivery of 3D video services which require high bit rates, low latencies, and the feasibility to deploy adaptive application layer adjustment, such as unequal error protection, that can tailor the error control and transmission parameters to the very different requirements exhibited by the different components of 3D video streams.

As there is such a great flexibility in the 3D video service over 4G network applications, the major challenge becomes how to efficiently utilize system resources supplied in different layers to maximize users' 3D viewing experience. From the perspective of the communication system, the allocation of system resources for a cross-layer designed framework is both constrained across different communication layers and constrained among users who are simultaneously sharing the same spectrum [3]. From the perspective of 3D video QoE, real-time 3D video transmission has further delay constraints for real-time playback, minimal requirement for binocular quality, more coding parameters with more decoding dependency to consider, and limited computation complexity on the mobile side. Moreover, the allocation of system resources should be conducted dynamically to reflect the time-varying characteristics of the channel condition and the time-heterogeneity of the video source. To acquire up-to-date information on system resource availability, the resource allocator often accrues extra computation costs and communication overhead depending on the accuracy and frequency. A formal way to resolve the resource allocation problem is to formulate a 3D over 4G network as a cross-layer optimization problem to maximize user QoE subject to system QoS constraints. The problem we often encountered is how to develop a good model to link/map between system QoS and user QoE. Also, we often have to deal with resources having both continuous and integer-valued parameters. The overall optimization problem may also have nonlinear or/and nonconvex objective functions/constraints, and many local optima may exist in the feasible range. Thus, obtaining the optimal solution is often NP hard. How to choose the parameter sets in different layers as the optimization search space and how to develop fast solvers to attain optimal/suboptimal values in real time remain challenging issues.

As there are more and more applications with high bandwidth requirement, and these services increase the demands on the 4G networks. A growing tendency toward increasing network capacity is to include architectures that tend to reduce the range of the radio links, thus improving the quality of the link (data rate and reliability) and increasing spatial reuse. In WiMAX, the solution is proposed by adopting the use of multi-hop networks. The other solution with an analogous rationale is to introduce the use of femtocells [4]. Femtocells are base stations, typically installed at a home or office, which operate with low power within a licensed spectrum to service a reduced number of users. Work in femtocells was initiated for 3G networks and it focused on addressing the main challenges when

implementing this technology: interference mitigation, management ease of configuration, and integration with the macrocell network. The progress achieved with 3G networks is being carried over to 4G systems, for both LTE and WiMAX. With the popularity of femtocells, 3D streaming over femtocells (or as the first/last mile) will become an important service. Therefore, there is a strong need to study how to efficiently allocate resources and conduct rate control.

In the multiuser video communications scenario, a single base station may serve several users, receiving the same 3D video program but requesting different views in free-viewpoint systems or multi-view video plus depth systems. As the information received by different users is highly correlated, it is not efficient to stream the video in a simulcast fashion. A more efficient way to utilize the communication resource is to jointly consider the correlation among all users' received 3D video programs and send only a subset of the video streams corresponding to a selection of views. The views actually transmitted are chosen in such a way that they can be used to synthesize the intermediate views. How to maximize all users' QoE by selecting the representative views and choosing the video encoding parameters and network parameters to meet different users' viewing preferences under different channel conditions becomes an important issue.

Most of the approaches in a 3D video communications pipeline take advantage of intrinsic correlation among different views. In networks of many users receiving 3D video streams, the view being received by one user could serve other users by providing the same view, if needed, or a view that could be used to synthesize a virtual view. This scenario could arise in multimedia social networks based on video streaming through peer-to-peer (P2P) network architecture. Some recent publications have studied techniques that address incentive mechanisms in multimedia live streaming P2P networks to enable the cooperation of users to establish a distributed, scalable, and robust platform. These techniques, nevertheless, fall into an area of incipient research for which still more work is needed. In the future, the properties of 3D video make it an application area that could benefit from techniques to incentivize the collaboration between users.

The features and quality offered by 4G systems result in an ecosystem where it is expected that users' equipment will present vastly heterogeneous capabilities. In such a background, scalability and universal 3D access become rich fields with plenty of potential and opportunities. With the introduction of 3D video services – in addition to the traditional spatial, temporal, and quality scalabilities – video streams will need to offer scalability in the number of views. For this to be realizable, it will be necessary to design algorithms that select views in a hierarchy consistent with scalability properties and that are able to scale up in a number of views by interpolation procedures. Considering scalability from the bandwidth perspective, it is also important to realize the challenges for 3D streaming services over low bit-rate network. Under a strict bit-rate budget constraint, the coding artifacts (e.g., blocking/ringing artifacts) due to a higher compression ratio become much more severe and degrade the binocular quality. Furthermore, channel resources for error protection are limited such that channel-induced distortion could further decrease the final 3D viewing experience. A joint rate control and error protection mechanism should be carefully designed to preserve/maximize objects with critical binocular quality (such as foreground objects) to remedy the coding artifacts and channel error. On the other hand, in order to achieve universal 3D access, 3D video analysis and abstraction techniques will attract more attention, as well as the transmission of 3D abstract data.

The development of distributed source coding paves the way for 3D video transmission over wireless network. Distributed source coding solution tries to resolve the problem of lossy source compression with side information. When applying this concept to video coding, this technique can be summarized as transmitting both a coarse description of the video source and extra data that completes the representation of the source, which is compressed using a distributed source coding technique (also known as Wiener–Ziv coding). The coarse description contains side information that is used to decode the extra data and obtain a representation of the reconstructed video. The main property exploited by the distributed video coding system is the correlation between the distributed source-coded data and the side information. The coarse description of the video can be either a highly compressed frame or an intra-predictive frame. In the latter case, the combination of the distributed source-coded data with the side information is able to recover the time evolution of the video sequence. Distributed video coding is a technique of interest for wireless video transmission because there is a duality between the distributed source-coded data and error correcting redundancy that results in an inherent resiliency for the compressed video stream.

We can extend the principle of distributed video coding to multi-view 3D video, since we can exploit the redundancies already present in mono-view video and add the new ones in multi-view video. Taking the simplest scenario consisting of two views, we can encode one view using distributed video coding and use the second view as side information. The other way to construct such a system is to deploy the distributed video coding for both views and use the implicit correlation between the two views to extract their time-dependent difference as side information. For a more generic setting involving more than two views, the multi-view video structure can be exploited by generating the side information from a combination of inter-view texture correlation and time-dependent motion correlation. Owing to the distributed nature, it is possible to combine the distributed 3D video coding with the use of relay nodes enabled with cooperative communications. Such a combination of distributed video coding and cooperative communications sets up a flexible framework that can be applied in a variety of ways. For example, different types of data in the multi-view video (or even V+D) can be via different channels/paths. Even more, if the relay is equipped with high computation ability, it can perform different application layer processing, such as transcoding/video post-processing/error concealment/view synthesis, to facilitate the 3D visual communication.

We often rely on the objective measurement, such as throughput, goodput, and mean-squared-error (MSE), to evaluate the system resource utilization from communication layer to 3D video application layer. One of the reasons is that the selected objective measurement simplifies the problem formulation by excluding the highly nonlinear HVS factors and the optimal solutions exist in the formulated linear/nonlinear continuous/integer optimization problem. However, the final 3D video quality is evaluated by the human eyes; and the objective measurement does not always align with what human beings perceive. In other words, understanding the 3D human vision system and quantifying the QoE becomes extremely important. More specifically, we need to find the critical features and statistics which affect the 3D QoE and an effective objective measurement for 3D QoE that reflects subjective measurement. It is also important to have a quantitative measurement mechanism to evaluate the impact of distortion caused in each stage

of the 3D communication pipeline to the end-to-end 3D QoE. Having those QoE metric will enables the QoE-based optimized framework for 3D visual communications.

References

1. D. Astely, E. Dahlman, A. Furuskar, Y. Jading Y, M. Lindstrom, and S. Parkvall, "LTE: the evolution of mobile broadband," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 44–51, April 2009.
2. S. Ahmadi, "An overview of next-generation mobile WiMAX technology." *IEEE Communications Magazine*, vol.47, no. 6, pp. 84–98, June 2009.
3. G.-M. Su, Z. Han, M. Wu and K. J. R. Liu, "Multiuser cross-layer resource allocation for video transmission over wireless networks," *IEEE Network Magazine, Special Issue on Multimedia over Broadband Wireless Networks* 2006; vol. 20, no. 2, pp. 21–27.
4. V. Chandrasekhar, J. Andrews, A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, September 2008.