MULSEMEDIA COMMUNICATION RESEARCH CHALLENGES FOR METAVERSE IN 6G WIRELESS SYSTEMS

Ian F. Akyildiz[1], Hongzhi Guo[2], Rui Dai[3], Wolfgang Gerstacker[4]

[1]Truva Inc., Alpharetta, GA, 30022, USA
[2]School of Computing, University of Nebraska-Lincoln, NE, 68588, USA
[3]Department of Electrical and Computer Engineering, University of Cincinnati, Cincinnati, OH, 45221, USA
[4]Institute for Digital Communications, Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstr. 7, D-91058 Erlangen, Germany

NOTE: Corresponding author: Hongzhi Guo, hguo10@unl.edu

Abstract – Although humans have five basic senses, sight, hearing, touch, smell, and taste, most multimedia systems today only capture two of them, namely, sight and hearing. With the development of the metaverse and related technologies, there is a growing need for a more immersive media format that leverages all human senses. Multisensory media (Mulsemedia) that can stimulate multiple senses will play a critical role in the near future. This paper provides an overview of the history, background, use cases, existing research, devices, and standards of mulsemedia. Emerging mulsemedia technologies such as extended Reality (XR) and Holographic-Type Communication (HTC) are introduced. Additionally, the challenges in mulsemedia research from the perspective of wireless communication and networking are discussed. The potential of 6G wireless systems to address these challenges is highlighted, and several research directions that can advance mulsemedia communications are identified.

Keywords – 6G wireless systems, augmented reality, extended reality, holographic-type communication, metaverse, mixed reality, mulsemedia, virtual reality

1. INTRODUCTION

The past several decades have seen significant progress in the development of multimedia communication and networking. Although the quality of video and audio has greatly improved, this has come at the cost of streaming high volumes of data through both wireless and wired networks. In recent years, the emergence of extended Reality (XR), digital twins, digital currency, and Holographic-Type Communication (HTC) [1] has led to the development of the metaverse. However, the multimedia used in the metaverse only stimulates sight and hearing, neglecting three other fundamental human senses: touch, smell, and taste. For the metaverse to become a reality where users can live and work as if they were in the real world, all five human senses must be addressed and stimulated to create a fully immersive experience. This calls for a new format of multimedia that can incorporate multiple human senses.

Mulsemedia research emerged in 2010 [2]–[4] and it stands for media that includes three or more human senses [5]. Although this concept is relatively new, the employment of humans' senses beyond sight and hearing in film theaters started about one century ago. The invention of television in 1927 attracted many film customers. Film theaters started to come up with novel technologies to bring more customers back. The first film that adopted extra human senses is Mein Traum in the 1940s, which uses a so called "smell-o-drama" to release odors into the theater. However, this technology was not widely adopted due to a lack of mature support. Around 1960, more attempts appeared in film theaters, including the famous "AromaRama" and "Smell-O-Vision" which release scents through different systems into theaters. Also, vibration was employed beneath seats. After this, the first 4D film was screened in 1984, incorporating multiple humans' senses, such as vibration and scent. Nowadays, the 4DX format provides richer human senses for 4D films. Moreover, significant efforts were also spent to create high-quality media for other human senses besides sight and hearing. For example, digital scent technology has been available for a long time and some devices can be integrated with modern XR headsets. A brief summary of the mulsemedia development milestones is given in Fig. 1.

In the past decade, researchers have developed various mulsemedia prototypes [6]–[9] which use a range of senses, including scent, light, moisture, wind, haptic, and kinesthetic effects. To create mulsemedia content, annotators manually add information about wind speed, light intensity, and scent to videos, which are then played together at the client side. These pioneering works have proven the feasibility of mulsemedia communications,

Year	Media/Technology	Name	Sense	Details
1940	Film	Mein Traum	Scent	"smell-o-drama", released 32 different odors into the theater.
1959	Film	The Tingler	Vibration	"percepto", buzzers underneath seats
1959	Film	Behind the Great Wall	Scent	"AromaRama", sends scents through air conditioning system into the theater.
1960	Film	Scent of Mystery	Scent	"Smell-O-Vision", pumps scents out of vents beneath the seats.
1984	Film	The Sensorium	Vibration/Scent	The first 4D film with bodysonic seats and Scent-a-Vision
1999	Digital scent	iSmell	Scent	Controlled by computers and generates various scents.
2009	4D film format	4DX	Multiple	Provides multisensory effects, including motion-seats, wind, light, scents, etc.
2020	Digital scent	OVR Technology	Scent	Scent devices that can be connected to XR headsets to generate scents.

Fig. 1 – A brief summary of mulsemedia development milestones.

and results have shown that mulsemedia can significantly improve the Quality-of-Experience (QoE) [10].

The metaverse [11] and the 6G wireless systems [12] have significantly changed the development of mulsemedia. First, the metaverse aims to provide seamless connections between the physical world and the virtual world, which demands truly immersive user experiences that can utilize more human senses beyond sight and hearing. This is the major driving application of mulsemedia communications. Second, the 6G wireless systems will provide ubiquitous connectivity with high data rates and low latency which enable mobile mulsemedia communications. Existing mulsemedia technologies are mainly used in 4D theaters due to complex sensing and display systems and high requirements on communication system performance. As a result, users have limited access to mulsemedia. The 6G wireless systems can dramatically change this situation by providing high-performance wireless communication for mulsemedia sensors, actuators and displays which can be embedded in computers and smart devices.

It is important to understand the motivation of mulsemedia research, review the essential system blocks of mulsemedia communication, and identify the research challenges in the era of metaverse and 6G wireless systems. In this paper, we provide a comprehensive review of the past and the current status of mulsemedia communication, upon which we discuss its potential use cases and research challenges. In Section 2, we introduce the background of the metaverse which demands immersive mulsemedia communication, and we give an overview of the role of 6G wireless systems in the metaverse and mulsemedia communications. We also point out the relations between emerging technologies, such as HTC and XR which includes Augmented Reality (AR), Mixed Reality (MR), and Virtual Reality (VR). After that, we present the use cases of mulsemedia communications in Section 3. The mulsemedia communication system architecture and the technologies utilized at the sources, networks, and destinations are introduced

in Section 4. In Section 5, we point out potential research opportunities in 6G wireless systems that can facilitate the development of mulsemedia communications. Finally, this paper is concluded in Section 6.

2. BACKGROUND: METAVERSE AND 6G WIRELESS SYSTEMS

In this section, we introduce the background of mulsemedia, including its important role in the metaverse and the enablers of XR and HTC supported by the 6G wireless systems.

2.1 Metaverse

The metaverse is defined as a hypothetical synthetic environment which is linked to the physical world [13]. Users can interact with each other in the virtual environment provided by the metaverse [14]. The metaverse is anticipated to become the next iteration of the Internet, consisting of immersive and hyperreal 3D digital virtual content. Technologies such as digital twins, avatars, and holograms enable the recreation and enrichment of the physical world in the metaverse. In addition to replicating physical environments, the metaverse allows for the creation of virtual information, objects, avatars, and environments that do not exist in the physical world. The integration of real and virtual content in the metaverse can greatly enhance user capabilities for work, entertainment, education, healthcare, and manufacturing [15].

The metaverse will come in various formats, ranging from simple computer games to complex digital twins, enabling a plethora of novel applications, such as entertainment, gaming, industrial design and simulation, and more. Cryptocurrency also plays an important role in the metaverse. Virtual currencies and assets based on Non-Fungible Tokens (NFTs), blockchain, and Web 3.0 enable social economics in virtual worlds with high-security levels. Moreover, blockchain and Web 3.0 provide fully distributed services and protect users' digital
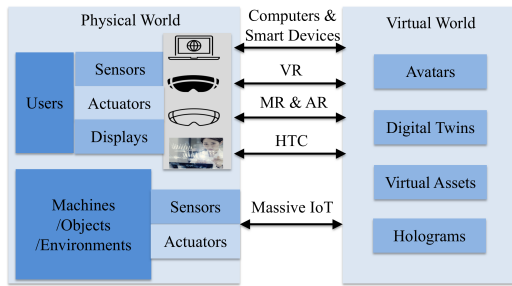
Fig. 2 – The gateways that connect the physical world and the virtual
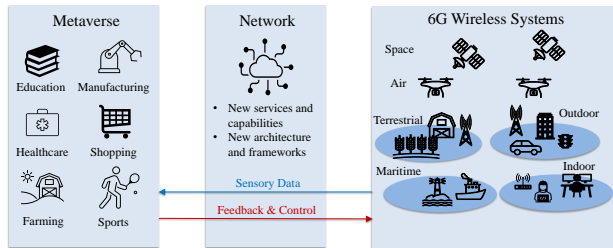

Fig. 3 – Illustration of the metaverse and 6G and beyond wireless systems.

identities.

Since the metaverse focuses on social interactions, ethical issues need to be considered in its development. Users must be accountable for their activities in the metaverse [16].

In order to access metaverse content, users require various devices such as smartphones, tablets, computers, Head-Mounted Displays (HMDs), light field displays, and others, as shown in Fig. 2. Unlike existing multimedia, high-quality metaverse content is created in formats such as 360-degree videos, holograms, and mulsemedia, among others. These new media formats represent complex information and require high data rates, often as large as several gigabits per second [1]. Existing multimedia streaming protocols may experience issues such as network congestion, storage overload, and long latency, which can drastically reduce the Quality of Experience (QoE) for users. Additionally, content creation for the metaverse relies on the Internet of Everything (IoE), which remains a challenge in 5G wireless systems. While the Internet of Things (IoT) only networks physical smart devices, the IoE networks people, things, data, and processes to provide intelligence and cognition across networked environments. As a result, the IoE demands more comprehensive and ubiquitous network connectivity and intelligence than the IoT.

Generally, wireless systems for the metaverse mainly require the following two types of communications, as shown in Fig. 2 and Fig. 3.

- Human and Metaverse Communication (HMC). As shown at the top of Fig. 2, the communication between users and the metaverse involves specialized devices, including HTC displays, sensors, XR HMDs, legacy computers, and smart devices, that

can be used to view and interact with the metaverse. These devices have sensing units that track users' motions and gather input data for computing in the metaverse. The computing results are sent back to users' devices for display and actuation.

- Machine and Metaverse Communication (MMC). Active and passive sensors can monitor physical objects, machines, and environments. The aggregated data can be used to develop hyperreal virtual models in the metaverse. Additionally, the metaverse can use actuators to interact with and change the physical world to meet users' request. The communication between the metaverse and sensors and actuators can use 6G wireless systems that support IoE.

It is important to note that HMC and MMC are bidirectional, reflecting the interactions between the physical and virtual worlds. For human-to-human and machine-to-machine communications, these interactions can take place in the physical world without the need for the metaverse, or they can use the metaverse as a gateway to become HMC or MMC. Mulsemedia communication is utilized to improve the QoE, and it is used for HMC.

Currently, the most popular tool for accessing the metaverse is the XR HMD [17]. However, some users experience frustration, eye strain, and sickness from prolonged use due to the large weight and low resolution of HMDs [18]. While working and living in the metaverse provides various choices and support that are not available in reality, existing HMDs are not convenient and can even create anxiety for some users, according to a recent study [18]. The low-quality of metaverse content, constrained by communication, computing, networking, and sensing capabilities, is the primary cause. Moreover, the HMDs integrate many electronic modules that have non-negligible weights, such as the battery which determines the operation time of wireless HMDs.

2.2 6G wireless systems

Users of the metaverse expect hyperreal and immersive virtual content that are compute and data-intensive [19]. Interaction with other users or virtual content in the metaverse demands ultra-low end-to-end latency and ultra-high reliability to avoid causing discomfort, such as dizziness, to users. As a result, the metaverse requires ultra-high data rates, ultra-high reliability, and ultra-low latency communications with computation capabilities in both core and edge networks. While 5G wireless systems have started the creation and development of the metaverse, providing high QoE services for the metaverse remains a challenge. 5G wireless systems are equipped with a range of services that can support metaverse applications, especially the enhanced Mobile Broadband (eMBB) and Ultra-Reliable Low Latency Communication (URLLC). eMBB provides high data rates and the URLLC prioritizes high reliability
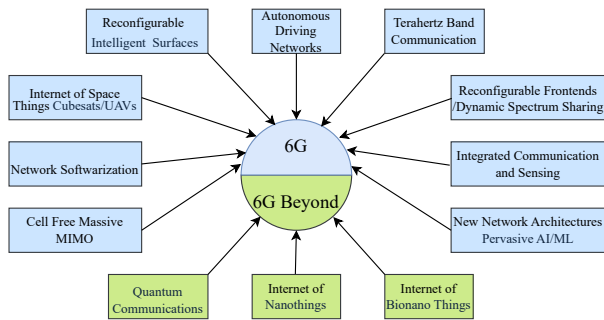
Fig. 4 – 6G and beyond technologies [12].

and low latency. However, the metaverse demands technologies that can integrate both of them simultaneously. It also requires various other services that are not available in 5G wireless systems, such as integrated communication and sensing and the deterministic networking with bounded end-to-end latency [1], [20].

The 6G wireless system [12], [20] is anticipated to provide ubiquitous, high-speed, low-latency, and ultra-reliable connectivity simultaneously, which is the convergence of the eMBB, URLLC, and other services of 5G wireless systems. These innovative services will be made possible by advancements in all-spectrum communications, including mmWave communications, terahertz communications, optical communications, joint communication and sensing, space-terrestrial integrated networks, cubesat networks, deterministic networks, mobile edge computing, artificial intelligence, and other technologies, as shown in Fig. 4. The physical world will be fully interconnected from space to air, terrestrial, maritime, indoor, and underground environments via 6G wireless systems, as shown in Fig. 3. This will provide global coverage and seamless connection between the virtual world and the physical world. The cloud and edge networks, equipped with new architectures and frameworks, can facilitate the delivery of new services and capabilities to support metaverse applications. The QoE of the metaverse is expected to be significantly improved by leveraging 6G wireless technologies in communications, networking, caching, and computation, and using flexible and lightweight wireless devices.

2.3 XR and HTC

XR is a type of spatial computing technology that enables users to observe and interact with virtual worlds [21]. According to Milgram and Kishino's Reality-Virtuality Continuum [22], XR technologies can be categorized based on their virtual content percentage and complexity. Augmented Reality (AR) only contains simple virtual content, with the majority of observations still being related to the real physical environment. Virtual Reality (VR) has full virtual content, with observations not related to the real physical environment. Mixed Reality (MR) combines both real and virtual content, with the virtual content often being interactive and dy-

namic. While there is overlap between MR and AR, MR is a broader concept that includes more interactive and richer virtual content.

Mulsemedia has been recently introduced to XR to provide more immersive user experiences [23]. XR devices can display and output various mulsemedia, including 360-degree videos, haptics, and scents. While smart devices like smartphones and tablets can be used to access XR content, they cannot provide highly immersive experiences as users must look at the screen. High-end AR and MR devices use optical see-through head-mounted displays (HMDs) to seamlessly integrate virtual content with the physical environment. Conversely, VR uses HMDs that completely block the users' view of the physical environment, and the displayed content is fully virtual. XR HMDs integrate various sensing, computing, communication, storage, and power units, including head tracking, hand tracking, eye tracking, proximity, and pressure sensors. These sensors enable XR devices to understand users' behavior and input. Mulsemedia sensors and displays can be embedded in XR HMDs without significant modifications. However, XR HMDs have limited computing, storage, and energy due to size and weight constraints. As a result, they usually need to pair with a server, which can be a cell phone, tablet, computer, or remote edge/cloud server, to offload computation tasks and obtain extra storage.

HTC is characterized by the delivery of holograms and other mulsemedia through wireless and wired networks [1]. A hologram is a recording of light field that preserves the original depth and parallax of real 3D objects. Hologram videos could provide 6 Degrees of Freedom (DoF) immersive viewing experiences, and users can freely move forward/backward (surging), up/down (heaving), or left/right (swaying) to select their favorite viewing angle of a 3D scene [24]. This advances the user experience over common VR applications, as VR video users are unable to move forward or backward freely. Furthermore, HTC systems may collect mulsemedia data using various sensors at the source and then regenerate the environment using various actuators, displays, and speakers at the destination. Users can select the senses depending on the available hardware and privacy and security issues.

A generic HTC system consists of the source, the destination, and the network [1]. The source uses various sensors to capture holographic content, synchronizes multisensory data, encodes the holographic data, and follows HTC networking protocols to send data packets. The HTC networks deliver source data with guaranteed performance in terms of bandwidth, latency, reliability, etc, which are defined by the HTC-enabled use cases. The destination receives and renders data for display, utilizes various actuators to regenerate the environment at the source, performs synchronization among multisensory media, and provides useful feedback to the source and the network if necessary. Users can observe truly 3D holograms with naked eyes using light field displays.
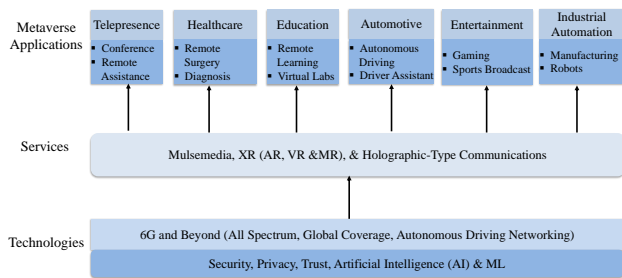
Fig. 5 – Metaverse applications, accessing technologies, and enabling technologies.
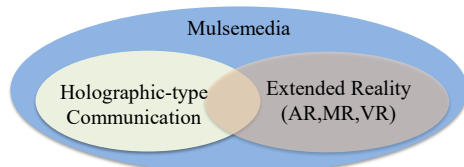

Fig. 6 – Metaverse and immersive mulsemedia technologies.

Together with other mulsemedia content, HTC can provide a high quality of experience. Typical HTC use cases can be found in [1].

The roles of mulsemedia communication and the associated XR and HTC technologies of the metaverse are shown in Fig. 5. The three key aspects of the metaverse are its applications, services (mulsemedia, XR, and HTC), and supporting technologies, especially the 6G wireless systems, as shown in Fig. 4 (more details of the 6G and beyond technologies can be found in [12]). Metaverse applications and users create the virtual worlds with social and economic activities, and XR and HTC are the key accessing technologies used to enter these virtual worlds. The relation between the metaverse, mulsemedia, XR, and HTC is shown in Fig. 6. Mulsemedia communication is a broad concept that includes all media that stimulate multiple human senses. XR and HTC are technologies that can deliver mulsemedia content. Mulsemedia can also be recorded and displayed using other traditional sensors and displays with reduced QoE. There are many ways to display holograms where XR HMD is one of these. XR HMDs can show different views, which can create 3D effects to display holograms. As a result, there is an intersection between XR and HTC.

3. USE CASES

In this paper, we explore four crucial use cases for mulsemedia communication in the fields of education and training, healthcare, retail, and entertainment. These domains are essential in our daily lives and provide an excellent opportunity to develop new mulsemedia communication theories, technologies, and business models. The selected use cases are representative of a broad spectrum of applications and demonstrate the varied potential applications of mulsemedia communication; however, they do not encompass the full scope of possible scenarios.

3.1 Education and training

Combining video and audio with sensory signals such as scent, haptic feedback, light, temperature, and more can create immersive learning experiences. For instance, students can touch, smell, and see flowers in biology and art classes using mulsemedia. This multisensory approach can enhance students' understanding and promote their learning experiences in various subjects, such as geology, where they can experience different weather patterns and conditions worldwide. Furthermore, mulsemedia can be a valuable tool for training professionals, such as first responders, miners, and farmers, as it enables them to recognize and fully grasp critical incidents and respond effectively. Mulsemedia can simulate the scents, temperature changes, and flashlights that professionals may encounter, making training more realistic and engaging.

3.2 Healthcare

Telehealth technology uses computer networks to connect healthcare providers with patients, improving healthcare accessibility and enabling doctors to treat contagious diseases without being in the same location as their patients. While traditional multimedia can only help to provide a basic diagnosis, patients must answer many related questions and describe their feelings. Mulsemedia communication can significantly increase diagnostic accuracy by allowing doctors to leverage more sensory information, such as haptic and temperature signals. Additionally, remote surgeries can be performed by doctors using mulsemedia communications to control robots or robotic arms/hands using haptic signals.

3.3 Retail

The retail industry has traditionally relied on strategies such as sending samples or using multimedia-based advertisements. However, customers often have to return goods because they lack sufficient information, which can only be obtained after receiving them. Mulsemedia offers a solution by providing a comprehensive description of a product through various senses. By using mulsemedia-based descriptions, customers can accurately experience the material, smell, and condition of goods before making an online purchase. As a result, this can increase customer satisfaction and provide greater success and revenues for retailers.

3.4 Entertainment and social interactions

Mulsemedia technology can be integrated into various existing smart devices for entertainment and social interactions. Mulsemedia can be used to develop truly immersive entertainment environments by stimulating
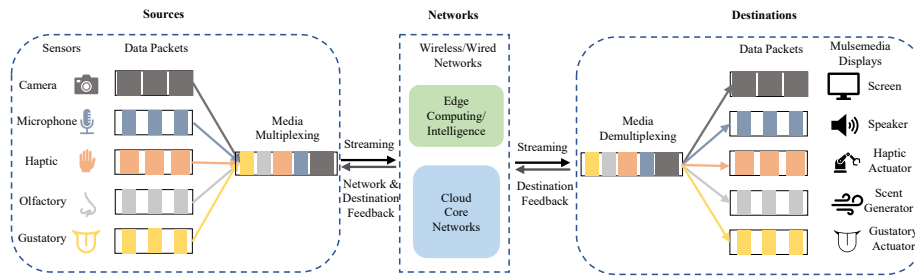
Fig. 7 – Mulsemedia communication system with sources, networks, and destinations. Multisensory content is multiplexed at the sources and demultiplexed at the destinations.

more human senses. In 2011, Samsung collaborated with the University of California at San Diego to develop a smart TV that can generate over 10,000 different odors. Additionally, smartphones, tablets, smartwatches, and other smart devices can process and display mulsemedia and allow users to interact with each other using various sensory media. This technology can be used to augment existing videos and pictures with additional senses, providing more immersive experiences. For instance, instead of sending simple videos, audios, or text messages, we can send friends various sensory messages, e.g., scent of followers.

4. MULSEMEDIA COMMUNICATION SYSTEMS

Existing multimedia communication schemes can be broadly divided into two categories, namely, Video-on-Demand (VoD) and Live Streaming (LS). The VoD content is prerecorded and streamed to users based on requests, while LS content is recorded and immediately sent to users. LS is more challenging than VoD since only limited time is available to record and preprocess multimedia content. Similarly, mulsemedia communications comprise of Senses-on-Demand (SoD) and LS, where "video" is replaced by "senses".

Existing mulsemedia communication systems are mainly SoD; LS remains a challenge. Although they require different streaming technologies, SoD and LS share the fundamental system units. Next, we introduce the three major components of a mulsemedia communication system, namely, the sources, the networks, and the destinations, as shown in Fig. 7.

4.1 Source

The source of SoD can be a server where mulsemedia is stored, while the source of LS can come from various devices such as computers, smart devices, sensory booths, or XR headsets. There are two methods of creating mulsemedia:

- *Annotation*. Additional senses can be added to existing multimedia files using annotation tools, such as PlaySEM [25]. Authors typically perform the annotation based on personal judgment. For instance,

when a video shows a flower, the author can annotate a smell based on previous experiences or knowledge, which may not necessarily be the exact scent of the flower.

- *Real-time recording*. Multiple senses are captured using various sensors simultaneously, and the sensing data is processed and encoded to create mulsemedia. The quality of mulsemedia in this approach is dependent on the sampling rate and accuracy of the sensors. However, this method results in a larger data size and higher data rates for communication compared to the annotation approach.

As mentioned earlier, the annotation-based approach for creating mulsemedia is relatively straightforward, and existing standards like MPEG-V [6], [26], which will be discussed in Section 4.4, have already provided comprehensive solutions. Therefore, we will focus on the real-time recording-based approach in the following sections.

4.1.1 Sensors

While mulsemedia communication poses several challenges, off-the-shelf sensors can capture various senses [27]. These include cameras, microphones, light intensity sensors, moisture sensors, temperature sensors, and wind sensors. Emerging haptic, olfactory (related to smell), and gustatory (related to taste) sensors are also available, as shown in Fig. 8. In this context, all displays and actuators used to replay mulsemedia are called displays. These sensors and displays can be integrated with existing smart devices, computers, and XR HMDs, as shown in Fig. 9.

Aside from the five basic senses, human beings also have many other senses, such as skin senses of moisture, temperature, and pressure. It is challenging to integrate all available sensors in XR HMDs due to the space and weight constraints. It is anticipated that XR HMDs need to be used together with wearable sensors to offload its complexity, weight, and power consumption. Furthermore, the HTC user with light field displays do not need any wearable devices to obtain a high QoE. Thus, novel designs of sensors without direct contact with the user's body are required.

Sense	Name	Sensor/Display	Year	Description
Haptic	Touch from 3D Systems	Sensor & Display	2014	User can manipulate on-screen 3D objects and feel force feedback on their hands.
	MANUS Haptic VR Gloves	Sensor & Display	2014	Interact with virtual models and generate haptic feedback.
Olfactory	Madeleine "Smell Camera"	Sensor	2013	Record odors and generate scentography.
	OVR ION	Display	2020	Wearable wireless digital scent generator.
	Olorama	Display	2013	Stimulate the sense of smell using more than 200 different scents.
Gustatory	e-Tongue by Insent	Sensor	1993	A sensor system using biomimetic membrane to provide high sensitivity of gustatory signals.
	Digital Lollipop	Display	2016	Digitally stimulate gustation using electric currents on the human tongue.
	Norimaki Synthesizer	Display	2020	Using electrolytes to supply mixed five basic tastes and generate arbitrary taste to the user's tongue.

Fig. 8 – Mulsemedia sensors and displays. The sensors and displays are 3D systems [28], MANUS Prime X Haptic VR gloves [29], Madeleine Smell Camera [30], OVR ION [23], Olorama [31], e-Tongue [32], Digital Lollipop [33], and Norimaki Synthesizer[34].
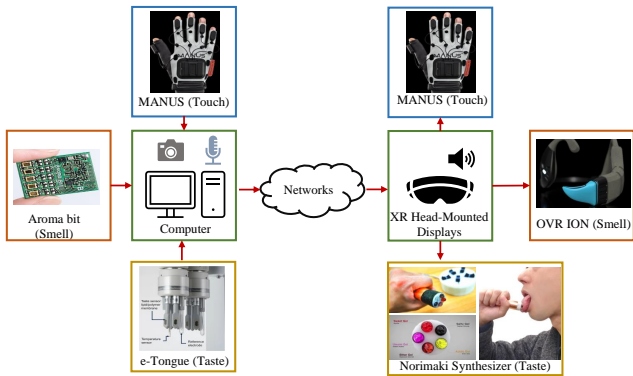

Fig. 9 – Mulsemedia communication system with a source (computer) and a destination (XR head-mounted display). The sensors and displays are e-Tongue [32], Aroma bit [35], MANUS Prime X Haptic VR gloves [29], OVR ION [23], and Norimaki Synthesizer[34].

Mulsemedia recording utilizes all available sensors to capture synchronized sensing media. While different sensors may have unique characteristics, the fundamental sensing mechanisms are similar. Sensors sample the sensing input signals periodically and compress and encode the resulting data. However, since sensing dynamics vary, for example, the environment temperature changes much slower than human motion, mulsemedia sensors require different sampling rates, data throughput, and sensitivity that depend on specific applications.

4.1.2 Mulsemedia encoding

Mulsemedia encoding can be performed by first encoding each sensory media and then multiplexing them for streaming, as illustrated in the left part of Fig. 7. In this way, mulsemedia can be created by slightly modifying existing multimedia encoding methods, rather than de-

signing new encoding systems from scratch. Furthermore, by increasing the sensing dimension and adding more sensory media to existing encoding systems, we can further enhance mulsemedia capabilities. Haptic, olfactory, and gustatory media are relatively new, and the corresponding encoding principles are similar to those used for other sensory media. For haptic signals, Apple uses AHAP format which is a JSON-like file. The IEEE 1918.1 Tactile Internet standard proposes haptic codecs including a kinesthetic codec and a tactile codec [36]. There are commercially available olfactory sensors and displays, but dedicated international encoding standards for olfactory media is currently not available. Gustatory media encoding faces a similar issue.

Fig. 10 shows the required data rates for each sensory media. As we can see, HTC requires the highest data rates due to the high complexity of holograms. Thus, besides inter-media encoding, the major intra-media encoding challenge lies in HTC. The current trend in presenting holograms is to use point cloud videos. A point cloud is a set of 3D points, where each point is associated with a geometry position and attribute information [37], such as color, reflectance, transparency, etc. 3D point clouds could be generated by synthesizing the videos captured simultaneously by multiple RGB cameras at different angles or multiple RGB cameras and depth cameras. For example, in the 8i Voxelized Surface Light Field (8iVSLF) dataset [38], the full body of a human subject was captured by 39 synchronized RGB cameras configured in either 12 or 13 rigs (or clusters), at a frame rate of 30 fps. The camera rigs were placed around the subject at approximately a few meters distance. Each cluster of cameras captured RGB and computed depth-from-stereo. The inputs from all data clusters were then fused into a 3D surface.

HTC systems usually include compression solutions to reduce the bandwidth demand of raw point cloud videos. MPEG Video-based Point Cloud Compression (MPEG V-PCC) and MPEG Geometry-based Point Cloud Compression (MPEG G-PCC) are the major standards for point cloud compression [37]. MPEG V-PCC encodes 3D point clouds as a set of 2D videos leveraging existing 2D video compression standards and tools. It is well suited for dense sequences that, when projected to 2D, result in image-like continuous and smooth surfaces. MPEG G-PCC encodes the content directly in 3D space. As the current G-PCC standard has only defined intra-prediction but no temporal prediction, it is more suitable for sparse meshes and static scenes.

Recent encoding experiments with MPEG V-PCC have demonstrated its good compression performance, producing decent quality for test point cloud sequences with a bit rate of 10 - 25 Mbps [37]. However, the encoding process for point cloud videos is time-consuming. As reported in [39], for a set of point cloud test sequences over a 7-10s period, the average coding time using the MPEG TMC13 software took as long as 8.28 seconds. There is a need to reduce the processing latency of en-

		Data Rates (Mbps)	Latency (ms)	Packet Loss
Video	HTC	1,000,000	< 1	$< 10^{-7}$
	XR	1,540	< 8	$< 10^{-7}$
	Fair Experience	60	<30	$< 10^{-4}$
Audio	High-quality	0.5	<10	$< 10^{-5}$
	Fair Experience	0.1	<20	$< 10^{-3}$
Haptic	High-quality	10	<1	$< 10^{-7}$
	Fair Experience	0.1	<50	$< 10^{-5}$
Olfactory	Fair Experience	0.1	<100	$< 10^{-3}$
Gustatory	Fair Experience	0.1	<100	$< 10^{-3}$

Fig. 10 – Estimated required data rates, end-to-end latency, and packet loss rates for five basic sensory media [36], [46].

coding for highly interactive and live broadcasting use cases. At the destination side, billions of devices are already capable of decoding V-PCC content, based on the readily available hardware 2D video decoding solutions. However, it is still challenging to implement V-PCC decoders on mobile devices, as decoding a V-PCC bitstream requires the synchronization of three video decoder instances, which is not adequately supported by current mobile operating systems [40]. The decoder also performs additional computations to render the 3D content for display. Therefore, the latency of the decoding process should also be optimized for live streaming scenarios.

In addition to the MPG V-PCC and G-PCC compression standards, there are a few other studies that address the compression of 3D/point cloud. Draco [41] is a software library (by Google) that aims to compress 3D geometric meshes and point clouds to enhance the storage and transmission of 3D graphics. Draco continuously splits the point cloud from the center utilizing the concept of KD tree formation, while also modifying the axes on each direction, followed by the application of entropy encoding techniques to compress the data [42]. Deep learning-based point cloud compression methods utilize neural networks to learn the underlying patterns and structures in the point cloud data, enabling them to efficiently represent the data with fewer bits. These methods typically use autoencoder networks or convolutional networks to learn a compact representation of the input data, which can then be decoded to reconstruct the original point cloud [43]–[45]. These methods have shown promising results in terms of compression rate and reconstruction quality, making them a potential solution for efficient storage and transmission of 3D point cloud data. However, these methods also have their own challenges, such as the need for large amounts of training data and the computational complexity of the network architectures.

4.2 Communication and networking

The annotation-based mulsemedia can be streamed through existing wireless and wired networks since it only adds negligible annotation metadata. Existing mul-

timedia protocols can be modified to add more sensory media for mulsemedia streaming, which will be discussed in Section 4.4. The research efforts that aim to improve data rates and reduce end-to-end latency in Wi-Fi and 5G wireless systems can be leveraged for mulsemedia communication.

Over-The-Top (OTT) multimedia streaming services could be used to adapt the transmission of mulsemedia according to network conditions. Current OTT streaming solutions are based on either the Transmission Control Protocol (TCP) or User Datagram Protocol (UDP). Retransmission is used in TCP when a packet is lost or not successfully received, which will increase latency and bandwidth requirements. Therefore, TCP-based solutions are more suitable for low-interactive applications such as the streaming of prerecorded content. A popular TCP-based solution for streaming prerecorded videos is Dynamic Adaptive Streaming over HTTP (DASH). DASH-based solutions for point cloud video streaming have been proposed recently [24]. At the server side, point cloud videos are divided into tiles and encoded under different quality levels, which can be selectively transmitted to save bandwidth. A tile selection module at the client side calculates the tiles residing in the user's Field-of-View (FoV) and selects the proper quality level for each tile to maximize the viewing experience. A DASH-based mulsemedia communication testbed has been developed in [9], where the synchronization issue is addressed and users obtain a high level of satisfaction.

UDP-based protocols are in favor of lower latency at the expense of a decreased reliability, and therefore, they have the potential to support high-interactive or LS mulsemedia applications. Traditional UDP-based streaming solutions, such as Web Real-Time Communications (WebRTC), work well for streaming 2D videos, but they need more investigation and redesign to render them suitable for transmission of mulsemedia data. Furthermore, to meet users' reliability requirements, UDP-based protocols should be combined with loss recovery technologies such as error-resilient encoding at the source and error concealment at the destination.

The transmission of LS mulsemedia poses significant challenges due to the high volume of sensory data flows from various sensors such as haptic, olfactory, gustatory, and others. These data flows can be multiplexed at the source and demultiplexed at the destination to reduce the complexity of multisensory data synchronization, as shown in Fig. 7. However, each sensory data flow has its unique characteristics, and it is more efficient to consider the embedded data structure and spatial-temporal correlations when encoding the data. Some sensors can record the data and directly send it through the networks to corresponding displays at the destination, reducing the end-to-end latency by simplifying the encoding and decoding processes. However, this method also has its challenges. First, the synchronization of multiple data flows can be difficult due to the

variance in the end-to-end latency of the best effort delivery service. Second, it requires large buffers and coordination of the displays, leading to large end-to-end latency. In general, both architectures have limitations and advantages. A hybrid system that can support both multiplexed and independent multisensory data flows may find a balance.

4.3 Destination

The destination of mulsemedia communication plays a critical role in the overall system. It performs four main tasks to ensure the successful reproduction of mulsemedia content. First, it receives and decodes the mulsemedia data packets that are transmitted over the network. Decoding can introduce some latency depending on the compression ratio used during encoding. Additionally, in case of transmission errors, the destination may request retransmission or an attempt to correct or conceal the errors. Second, it synchronizes the multisensory data that arrive at different times. The destination must compensate for any variations in latency that may arise due to network delays or other factors. Third, the destination needs to coordinate the recreation of the mulsemedia using various displays, such as actuators, speakers, monitors, etc. This task involves optimizing the scheduling of display activities considering the unique characteristics of different devices. Finally, the destination can provide feedback to the source regarding the status of the network for adaptive transmission. By performing these tasks, the destination ensures that mulsemedia content is accurately and efficiently reproduced, leading to a high-quality user experience.

Due to the high volume of hologram data, recreating the hologram is the most complex task at the destination. There are two major types of displays that allow the viewing of holograms with naked eyes with a much better user experience than by using XR HMDs or smart devices: multiview volumetric display and light field display. The multiview volumetric display is a relatively small holographic display that can be used for mobile devices and holographic monitors. Usually, it can only support one user, and eye-tracking technologies can be used to adaptively render the holographic content. Light field displays can provide thousands of view angles and support multiple users simultaneously. Light field displays are anticipated to be widely used in future HTC and mulsemedia systems since they can provide a hyperreal user experience. To display real-time holograms of human subjects using a 1.5 m $\times$ 0.75 m display with 30 fps, a light field display requires an uncompressed data rate of 6.63 TBps [47]. Such a high data rate is beyond the transmission capacity of today's networks, which necessitates new designs of wireless communication systems to support mulsemedia communications with light field displays at the destination.

4.4 Standards

MPEG-V is an international standard for mulsemedia communication which consists of the following parts: architecture, control information, sensory information, virtual world object characteristics, data formats for interaction, common types and tools, and reference software and conformance [26]. Its main goal is to enhance interactions between the real and virtual worlds and support XR applications using mulsemedia. The sensory information part of the standard defines various senses, including light, temperature, wind, vibration, sprayer, scent, fog, color correction, motion, kinesthetic, and tactile. This sensory information is annotated as Sensory Effect Metadata (SEM) and streamed together with multimedia files as shown in Fig. 11. The SEM file provides detailed information about the multisensory media, including the sense type, intensity, duration, start time, and other attributes. The lower part of Fig. 11 shows an example of the sensory effect description in XML. At the destination, the SEM file is decoded and synchronized with the multimedia file.

The SEM created using MPEG-V has a negligible size which can be easily attached to existing multimedia formats. However, the multisensory effects defined in SEM are simple and only based on annotated descriptions. On the contrary, videos and audios are created using sampling which produces many samples and data packets in one second. Despite MPEG-V providing multisensory effects, it is challenging to achieve high QoE mulsemedia and realize LS of mulsemedia.

The inclusion of more sensory information in XR content is a growing trend in the multimedia communication field. The MPEG Immersive Video (MIV) standardization effort is an example of this, as it aims to incorporate haptic effects in videos with 6 DoF, providing users with a more immersive experience in the metaverse and other XR applications. Despite significant advancements in video and audio technologies, the development of other sensory media has been lagging behind. This presents a challenge for mulsemedia communication researchers to overcome. In the next section, we will explore some current research challenges and future opportunities in this field.

The IEEE 1918.1 Tactile Internet standard focuses on haptic signal transmission through the Internet [36]. The haptic signal can be sent together with videos and audios and thus the Tactile Internet can be considered as another mulsemedia application. The haptic data size depends on the DoF. For immersive virtual reality, the DoF ranges from 1 to 6, while for complex tele-operation the DoF can be as high as 100. The haptic codec consists of kinesthetic codec and tactile codec. The kinesthetic codec is developed for 3D position, velocity, force, and torque data, which considers two different scenarios, namely delay intolerant and delay tolerant. The tactile codec considers hardness, thermal conductivity, friction, microroughness, and macroroughness with re-
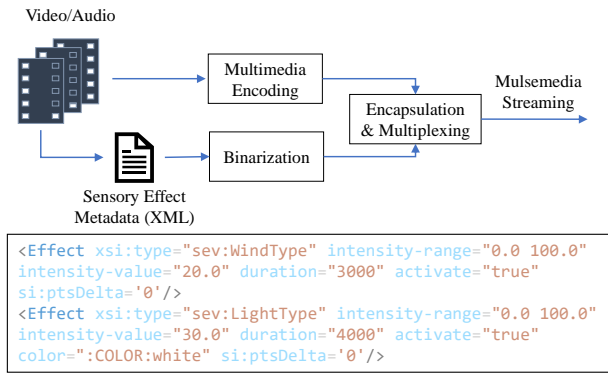
```
<Effect xsi:type="sev:WindType" intensity-range="0.0 100.0"
intensity-value="20.0" duration="3000" activate="true"
si:ptsDelta='0'/>
<Effect xsi:type="sev:LightType" intensity-range="0.0 100.0"
intensity-value="30.0" duration="4000" activate="true"
color=":COLOR:white" si:ptsDelta='0'/>
```

Fig. 11 – Mulsemedia transmission using MPEG-V. The lower part is an example of the sensory effect metadata which describes the effects of wind and light.
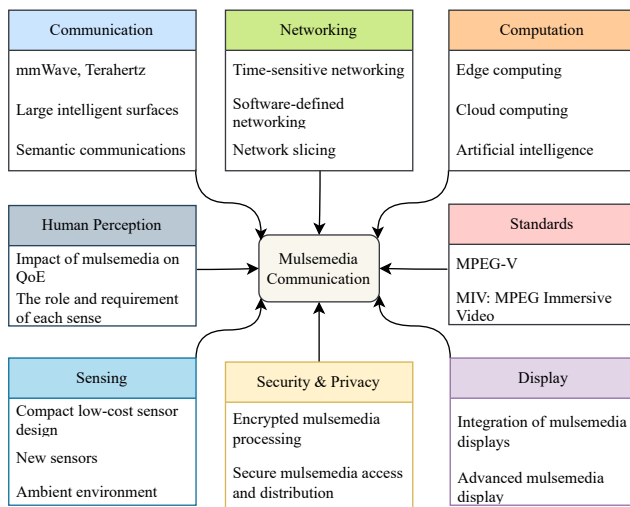

Fig. 12 – An overview of mulsemedia communication research areas and technologies.

laxed delay requirements.

5. RESEARCH CHALLENGES AND POTENTIAL SOLUTIONS

This section delves into the research challenges and potential solutions for mulsemedia communication. An overview of the research challenges and areas is given in Fig. 12. We consider XR and HTC are the technologies used for mulsemedia communications. The following technical requirements are essential for mulsemedia communications:

- Reliable high-speed, low-latency wireless communication to enable unrestricted movement. Existing mulsemedia communication systems use sensors and displays networked by cables, which significantly limit users' mobility. Some existing XR HMDs are also tethered by cables.

- Accurate high-fidelity wireless sensing technologies that can monitor and predict users' motion and render content adaptively. Integrated head, hand, and eye tracking sensors in existing XR HMDs are used, but the ultimate goal is to remove HMDs to allow users to play mulsemedia without wearing any devices. Wireless sensing of the motion of head, hand, body, and eyes can effectively free users.

- New Internet protocols and wireless networking technologies are needed to provide new services and capabilities for delivering mulsemedia content with different quality-of-service.

- Computing and content caching must move from the core cloud towards the network edge to reduce end-to-end latency and support highly interactive mulsemedia applications. Real-time solutions will require the consideration of processing tasks for mulsemedia content generation, compression, and rendering, which could be time-consuming.

- Computing and communication solutions have to be designed to boost user experience and leverage the characteristics of QoE for system optimization.

5G wireless systems are unable to meet the requirements due to the following shortcomings.

- **Data rates**. While 5G wireless systems promise data rates exceeding 1 Gbps, recent measurements have shown that the throughput can be as low as 100 Mbps, which is insufficient for mulsemedia applications [48], [49]. Although mmWave can provide a higher throughput of around 2 Gbps, the throughput is not stable and requires more intelligent management to maintain the desired performance [50]. As shown in Table 1, the experienced data rates are much smaller than the peak data rates.

- **End-to-end latency**. 5G wireless systems promise an end-to-end latency of 1 ms, but recent measurements have shown that the latency is much higher due to constraints in the wire-line networks [48]. Private networks with reserved resources can reduce the latency significantly, but it still remains higher than 1 ms [49].

- **Reliability**. 5G wireless systems achieve much higher reliability than 4G wireless systems [49]. For 5G standalone networks, the core network packet loss rate is on the order of 10^{-6}% and the radio access network packet loss rate is $3.28 \cdot 10^{-4}$%, as measured in [49]. However, for large size packets, which are common in mulsemedia applications, the packet loss rate increases significantly at the air interface to around 8.17% [49].

In addition to these limitations, 5G wireless systems have shortcomings with respect to energy efficiency, coverage, sensing, computing, and caching. On the contrary, 6G wireless systems have the potential to address

Table 1 – Comparison of 5G and 6G wireless systems

	5G	6G
Peak data rates	20 Gbps	1 Tbps
Experienced data rates	0.1 Gbps	1 Gbps
End-to-end latency	1 ms	0.1 ms
Reliability	10^{-5}	10^{-9}
Frequency bands	sub-6GHz, mmWave	sub-6GHz, mmWave, THz

most of these challenges. In the following, we will discuss the challenges to meeting mulsemedia communication requirements and their potential solutions.

5.1 Wireless communication

To facilitate mobility, wireless communication is a critical technology to connect mulsemedia sensors and actuators. In this subsection, we discuss the physical layer challenges and potential solutions. To achieve low latency and high data rates, terahertz communication using frequencies from 0.1 to 10 THz can provide data rates of several terabits per second (Tbps), which is an important advancement of 6G wireless systems [51]. Such a solution can be used to establish high-speed networks for mulsemedia sensors and displays. The range of communication can span several meters to tens of meters, depending on the location of mulsemedia devices. Accurate terahertz band channel models are necessary to enable adaptive wireless resource allocation. Additionally, the physical layer and data link layer communication protocols need to be reinvestigated and jointly designed by considering the specific mulsemedia applications. To prolong the operation time of wireless mulsemedia devices and avoid frequent recharging, new signal waveforms and media access technologies have to be developed in the terahertz band to improve energy efficiency.

Ubiquitous connectivity is required since users need to get access to mulsemedia content from any location. The space-air-ground-sea integrated networking is an essential technology to achieve this goal [15]. Space satellites and Unmanned Aerial Vehicles (UAVs) enable larger coverage that can provide wireless connectivity also in isolated areas without terrestrial wireless infrastructures.

For mulsemedia users with XR and HTC devices who have strong interactions with the virtual world, their activities can incur various body motions. Since terahertz wireless signals are highly directional, the alignment between terahertz transceivers is challenging due to the required precision. Also, the motion of the human body can affect the propagation of wireless signals. To this end, the Reconfigurable Intelligent Surfaces (RIS) can be leveraged to form a smart wireless environment to adaptively reflect and reroute wireless signals to en-

sure that the wireless link can be smoothly maintained [52][53].

Semantic communication will play an important role in mulsemedia communication. Instead of transmitting information bits, semantic communication uses machine learning/deep learning tools to extract underlying meaning in text, images, audio, and video [54]–[56]. Existing research has shown that semantic communication can be more reliable in the presence of noises and efficient in terms of delivering information [54], [55]. In learning-based semantic communications, typically different neural network modules related to source encoding, channel encoding, modulation, channel decoding and source decoding are jointly optimized according to a loss function tailored to the considered media format. In [57], a semantic communication approach is presented in which a predefined Binary Phase-Shift Keying (BPSK) constellation is employed for modulation. It should be noted that in most other semantic communication schemes proposed in the literature, impractical analog constellation symbols are generated and transmitted in order to avoid non-differentiable functions which cannot be trained with existing gradient-based algorithms. In [57], this problem is addressed by proposing a neural encoding network generating likelihoods of constellation points which is combined with a random coding-based modulator. The corresponding results demonstrate that existing digital modulation methods in semantic communications and neural network-based analog modulation, respectively, can be outperformed by the proposed scheme. More research in this direction is required in order to design efficient and practically relevant joint modulation and coding schemes for mulsemedia semantic communications with arbitrary discrete symbol constellations.

Although mulsemedia communication systems leverage multiple human senses, the generated multisensory media have strong semantic correlations because they represent the same perception of the physical world. Multimodal machine learning [58] is an extensively studied machine learning area that is strongly related to mulsemedia semantic communication. Existing work mainly studies the semantic correlations among text, audio, image, and video. By using deep generative models, we can generate an unobserved media format with one or several available media formats, e.g., from text and image to videos [59]. This approach can be extended to mulsemedia communication and generate multisensory media that cannot be easily captured by sensors.

5.2 Integrated communication and sensing

Wireless sensing can reduce the weight and power consumption of wearable mulsemedia devices, such as XR HMDs, by using external access points, base stations, or ambient wireless signals. Wireless sensing can enhance the recording and display quality of mulsemedia content. Motion sensing is essential for mulsemedia com-

munication systems, as it enables adaptive rendering of mulsemedia content. For example, in XR systems the communication data rate and virtual content can be optimally selected based on the FoV [60], which is enabled by FoV sensing. Also, prediction of a user's motion based on previous sensing knowledge can reduce the end-to-end latency. Machine learning can be used to predict the FoV based on the user's motion and preselect the optimal communication parameters.

6G wireless systems can leverage terahertz signals for motion sensing, as they have a high sensing resolution due to their short wavelength [51]. Terahertz beamforming using massive MIMO can track users and sense their motion with steerable beams. Terahertz sensing can also use active radar systems with dedicated transmitters and receivers or ambient systems that analyze and detect changes in the propagation and distribution of terahertz signals of the background environment. While terahertz radar has been used to classify gestures [61], its accuracy can still be improved.

It should be noted that sensing should be codesigned with communication in 6G wireless systems. Joint Communication and Sensing (JCAS) is an integral part of 6G networks [62]. The massive MIMO systems, RIS, mmWave signals, terahertz signals, and many other wireless technologies provide high resolution of sensing. The coexistence of sensing and communication poses new research problems. New algorithms and protocols are desirable for mulsemedia applications.

5.3 Mulsemedia streaming

Mulsemedia communication requires unprecedented high throughput (several Tbps) to meet the demands of multiple users. Such high throughput necessitates either a new network architecture design or improvements to the existing one. The New IP protocol introduced in [63] expands existing Internet protocols to add more services and mechanisms, such as adaptive packet dropping and guaranteed latency. Various network support technologies are essential for mulsemedia streaming, such as in-network caching, bandwidth saving, and adaptive streaming [64]. To achieve high data rates and low end-to-end latency, simply increasing wireless communication capacity is not effective. Existing studies of 5G wireless systems have shown that the wireline network is the bottleneck for reducing end-to-end latency [48]. The edge and core networks also need updating to provide new services, such as precision networks that can deliver data packets at predefined times with guaranteed variances. Additionally, edge caching can efficiently reduce network traffic. For example, caching frequently requested information such as the floorplan, store information, and product information in local edge servers for AR and MR users in a shopping mall can optimize storage and communication latency.

Besides general network upgrade, mulsemedia streaming can be specifically developed by adding additional

sensing media to existing multimedia streaming protocols. The hidden visual channel communication or hidden acoustic channel communication can be used to deliver mulsemedia [65]. In [65], the metadata is modulated using Chirp Spread Spectrum (CSS) in the format of acoustic signals. The Advanced Audio Coding (AAC) encodes the metadata together with audio signals. The destination decoder is carefully designed to extract the metadata without affecting audio signal quality. Similar techniques can be used to insert the metadata into visual channels for video encoding. The metadata can include the mulsemedia information such as the SEM in MPEG-V. Although this approach is low-cost and low-complexity, the amount of mulsemedia data that can be transmitted is limited. The metadata can be transmitted with 156.25 bps in [65]. It remains a challenge to support interactive mulsemedia applications.

5.4 Mulsemedia synchronization

In Fig. 10, we show the required data rates, latency, and packet loss rate for five basic sensory media. The high-quality haptic media can be used for teleoperations and the fair experience requirements are simple haptic signals that can be used for gaming with haptic controllers. The olfactory and gustatory requirements are suggested values based on other sensory media since the products are not widely available. As we can see, the bottleneck of mulsemedia data rates is video and all other media have much smaller data rates. However, it is not trivial to include more senses due to the complex synchronization problem [66], [67].

Mulsemedia communication requires strict synchronization of multiple sensory data flows, which includes (1) intra-media synchronization and (2) inter-media synchronization. The intra-media synchronization reorganizes received data packets of single sensory media, while the inter-media synchronization coordinates the display of multiple sensory media, as shown in Fig. 13. Similar problems have been already investigated in multimedia (audio and video) communication in the 1990s [68]–[71]. In particular, the lip-synch problem (jitter problem) was investigated by many researchers [68], [72]. However, in mulsemedia communication the number of sensory media to be synchronized are significantly increased and, accordingly, there are many more additional research challenges.

Note that if the buffering solution is used then the end-to-end latency can be affected which is a strict QoE in mulsemedia communication. First, the application can reorder packets after receiving and buffering them. However, this requires a sufficient buffer size and the application needs to be delay-tolerant. Another possible solution could be using traffic shaping solutions [73] such as leaky bucket and token bucket algorithms from existing literature. Despite improvements in communication data rates and network latency reduction in previous generations of wireless networks, to the best of
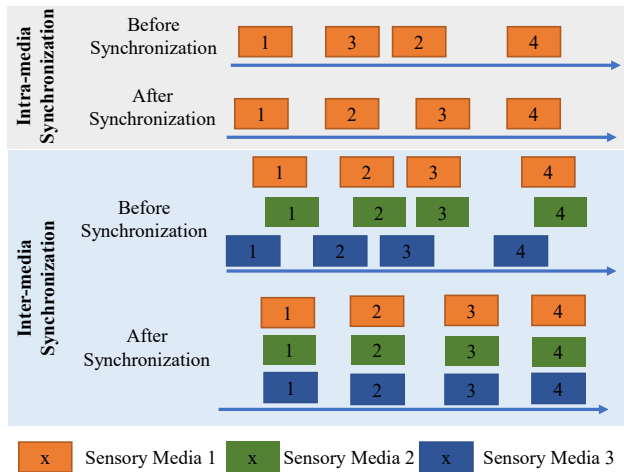
Fig. 13 – Illustration of intra-media and inter-media synchronization.

our knowledge there is currently no wireless technology capable of delivering data packets with bounded end-to-end latency and synchronized, coordinated data flows. This differs from the requirement of low latency, which would only guarantee that a large percentage of data packets meet the deadline with a certain number of outliers. As a result, new technologies are required for 6G wireless systems to meet the demands of mulsemedia communication.

5.5 Artificial intelligence

Besides semantic communication discussed in Section 5.1, artificial intelligence, especially basic machine learning and deep learning models, can be used in many other areas of mulsemedia communication systems to improve system performance. To enable widespread adoption of mulsemedia, it is essential to explore immersive video and hologram generation using more accessible devices such as smartphones. However, current technologies for generating and rendering immersive videos and holograms require numerous fixed camera sensors and sophisticated processing tools that are not widely available. Moreover, the generation, encoding/decoding, and rendering of 3D visual information involves several processing steps that can cause significant delays. Therefore, it is expected that mobile edge computing will play a crucial role in providing additional computation capability. Edge intelligence techniques [74], such as intelligent data compression and intelligent offloading of high-complexity tasks to the cloud, can help accomplish computational tasks on time. To ensure that end-to-end latency requirements of high-interactive applications are met, lightweight computation algorithms should be designed. It is also essential to jointly design computation and communication components, considering multiple sources of latency.

The recently developed Large Language Model (LLM) can also facilitate the design and commercialization of mulsemedia communication systems. It has been shown that LLM can be used to design chips [75] and this idea can be extended to communication systems design [76]. Multimodal large models can be trained on data collected from existing data communication systems which can be used to simplify the design and management of existing wireless and wired communication networks. Mulsemedia communication system designers can interact with data communication LLM and obtain suggestions and even solutions. Besides interactive data communication LLM, a dedicated Large Data Communication Model (LDCM) can be trained for autonomous driving networking, which can self-heal, self-diagonose, self-monitor, self-configure, and self-report. Mulsemedia communication users can send a request given required sensory media, bandwidth, end-to-end latency, etc. The LDCM will autonomously schedule resources and connect the source user with the destination. The research in zero touch networks, autonomous driving networks, and self driving networks has the potential to realize this vision [77], [78].

5.6 QoE modeling, management, and optimization

QoE measures the degree of delight or annoyance of users. It can be obtained by using subjective measures, where users rate their experience after using the service or application, and objective measures, where physiological signals such as accelerometers, video recordings, and biosignals are used to evaluate users' biofeedback and responses [10], [79]. A model of QoE is essential to design optimal mulsemedia communication systems. The QoE of mulsemedia can be modeled as a linear or nonlinear combination of each individual sense's QoE. The coefficients for this combination can be determined by machine learning models trained on data collected from system configurations and user ratings. Also, it is interesting to investigate the impact of physical layer parameters such as transmit power and power of interference from other users, and physical parameters influencing the behavior of the THz channel like the water vapor percentage, on the QoE of users. Such study has been conducted in [80] for VR users, indicating that using frequencies with low molecular absorption and the presence of a line-of-sight path are crucial. As mentioned in [80], a separate analysis should be performed for holographic/mulsemedia communications which have different QoE requirements. For example, five senses have to arrive synchronized, i.e., the instantaneous delay stemming from each sense is identical and extremely low. Furthermore, it would be of great interest to extend the model in [80] such that it reflects the dependency of QoE on the detailed waveform of transmission, and to employ such a tool for waveform optimization. With a QoE model, mulsemedia streaming can be optimally designed by optimizing communication systems, networking protocols, and available resources

and adaptively streaming mulsemedia.

6. CONCLUSION

Mulsemedia technology has the potential to provide users with highly immersive experiences that engage multiple senses beyond just sight and hearing. By transmitting multisensory content through both wired and wireless networks, mulsemedia communication represents an exciting and rapidly evolving field of study. In this paper, we provide an overview of the background, history, and essential encoding, decoding, and networking technologies that underpin this emerging field. We introduce the two major mulsemedia applications, namely, extended Reality (XR) and Holographic-Type Communication (HTC). Additionally, we highlight potential research directions for mulsemedia communication in the context of 6G wireless systems, which have the potential to further enhance the capabilities and performance of this exciting technology.

ACKNOWLEDGEMENT

REFERENCES

[1] I. F. Akyildiz and H. Guo, "Holographic-type Communication: A New Challenge for The Next Decade," *ITU Journal on Future and Evolving Technologies*, vol. 3, no. 2, pp. 421–442, 2022.

[2] G. Ghinea and O. Ademoye, "A user perspective of olfaction-enhanced mulsemedia," in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, 2010, pp. 277–280.

[3] R. Kannan, S. Balasundaram, and F. Andres, "The role of mulsemedia in digital content ecosystem design," in *Proceedings of the international conference on management of emergent digital ecosystems*, 2010, pp. 264–266.

[4] R. Kannan and F. Andres, "Digital library for mulsemedia content management," in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, 2010, pp. 275–276.

[5] G. Ghinea, C. Timmerer, W. Lin, and S. R. Gulliver, "Mulsemedia: State of the art, perspectives, and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 1s, pp. 1–23, 2014.

[6] M. Waltl, B. Rainer, C. Timmerer, and H. Hellwagner, "An end-to-end tool chain for Sensory Experience based on MPEG-V," *Signal Processing: Image Communication*, vol. 28, no. 2, pp. 136–150, 2013.

[7] D. Marfil, F. Boronat, J. González, and A. Sapena, "Integration of Multisensorial Effects in Synchronised Immersive Hybrid TV Scenarios," *IEEE Access*, vol. 10, pp. 79 071–79 089, 2022.

[8] A. Covaci, E. B. Saleme, G. Assres, I.-S. Comsa, R. Trestian, C. A. S. Santos, and G. Ghinea, "Multisensory 360 videos under varying resolution levels enhance presence," *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[9] T. Bi, A. Pichon, L. Zou, S. Chen, G. Ghinea, and G.-M. Muntean, "A DASH-based mulsemedia adaptive delivery solution," in *Proceedings of the 10th International Workshop on Immersive Mixed and Virtual Environment Systems*, 2018, pp. 1–6.

[10] Z. Yuan, S. Chen, G. Ghinea, and G.-M. Muntean, "User quality of experience of mulsemedia applications," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 1s, pp. 1–19, 2014.

[11] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Communications Surveys & Tutorials*, 2022.

[12] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE access*, vol. 8, pp. 133 995–134 030, 2020.

[13] L.-H. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, A. Kumar, C. Bermejo, and 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.

[14] Y. K. Dwivedi, L. Hughes, A. M. Baabdullah, S. Ribeiro-Navarrete, M. Giannakis, M. M. Al-Debei, D. Dennehy, B. Metri, D. Buhalis, C. M. Cheung, *et al.*, "Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 66, p. 102 542, 2022.

[15] F. Tang, X. Chen, M. Zhao, and N. Kato, "The Roadmap of Communication and Networking in 6G for the Metaverse," *IEEE Wireless Communications*, 2022.

[16] C. B. Fernandez and P. Hui, "Life, the metaverse and everything: An overview of privacy, ethics, and governance in metaverse," in *2022 IEEE 42nd International Conference on Distributed Computing Systems Workshops (ICDCSW)*, IEEE, 2022, pp. 272–277.

[17] I. F. Akyildiz and H. Guo, "Wireless Communication Research Challenges for Extended Reality (XR)," *ITU Journal on Future and Evolving Technologies*, vol. 3, no. 2, pp. 273–287, 2022.

[18] V. Biener, S. Kalamkar, N. Nouri, E. Ofek, M. Pahud, J. J. Dudley, J. Hu, P. O. Kristensson, M. Weerasinghe, K. Č. Pucihar, *et al.*, "Quantifying the effects of working in VR for one week," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 11, pp. 3810–3820, 2022.

[19] Y. Cai, J. Llorca, A. M. Tulino, and A. F. Molisch, "Joint compute-caching-communication control for online data-intensive service delivery," *arXiv preprint arXiv:2205.01944*, 2022.

[20] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, 2019.

[21] S. Shekhar, S. K. Feiner, and W. G. Aref, "Spatial computing," *Communications of the ACM*, vol. 59, no. 1, pp. 72–81, 2015.

[22] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, "Augmented reality: A class of displays on the reality-virtuality continuum," in *Telemanipulator and telepresence technologies*, SPIE, vol. 2351, 1994, pp. 282–292.

[23] "OVR ION." (), [Online]. Available: `https://ovrtechnology.com/`.

[24] Z. Liu, Q. Li, X. Chen, C. Wu, S. Ishihara, J. Li, and Y. Ji, "Point cloud video streaming: Challenges and solutions," *IEEE Network*, vol. 35, no. 5, pp. 202–209, 2021.

[25] E. B. Saleme and C. A. S. Santos, "PlaySEM: a platform for rendering MulSeMedia compatible with MPEG-V," in *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web*, 2015, pp. 145–148.

[26] K. Yoon, S.-K. Kim, J. J. Han, S. Han, and M. Preda, *MPEG-V: Bridging the virtual and real world*. Academic Press, 2015.

[27] Y. Sulema, "Mulsemedia vs. multimedia: State of the art and future trends," in *2016 International conference on systems, signals and image processing (IWSSIP)*, IEEE, 2016, pp. 1–5.

[28] "3D Systems." (), [Online]. Available: `https://www.3dsystems.com/`.

[29] "MANUS Prime X Haptic VR gloves." (), [Online]. Available: `https://www.manus-meta.com/products/prime-x-haptic`.

[30] D. Coldewey. "Snap 'n sniff? Madeleine camera captures scent, not light." (), [Online]. Available: `https://www.nbcnews.com/technolog/snap-n-sniff-madeleine-camera-captures-scent-not-light-6c10520866`.

[31] "Olorama Technology." (), [Online]. Available: `https://www.olorama.com/`.

[32] "Insent | Intelligent Sensor Technology, Inc." (), [Online]. Available: `https://www.insentjp.com/`.

[33] N. Ranasinghe and E. Y.-L. Do, "Digital lollipop: Studying electrical stimulation on the human tongue to simulate taste sensations," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 1, pp. 1–22, 2016.

[34] H. Miyashita, "Norimaki synthesizer: Taste display using ion electrophoresis in five gels," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–6.

[35] "Aroma Bit, Inc." (), [Online]. Available: `https://www.aromabit.com/en/`.

[36] O. Holland, E. Steinbach, R. V. Prasad, Q. Liu, Z. Dawy, A. Aijaz, N. Pappas, K. Chandra, V. S. Rao, S. Oteafy, *et al.*, "The IEEE 1918.1 "tactile internet" standards working group and its standards," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 256–279, 2019.

[37] D. Graziosi, O. Nakagami, S. Kuma, A. Zaghetto, T. Suzuki, and 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*, vol. 9, 2020.

[38] M. Krivokuca, P. A. Chou, and P. Savill, "8i voxelized surface light field (8iVSLF) dataset," *ISO/IEC JTC1/SC29/WG11 MPEG, input document m42914*, 2018.

[39] H. Liu, H. Yuan, Q. Liu, J. Hou, and J. Liu, "A comprehensive study and comparison of core technologies for MPEG 3-D point cloud compression," *IEEE Transactions on Broadcasting*, vol. 66, no. 3, pp. 701–717, 2019.

[40] S. Schwarz and M. Pesonen, "Real-time decoding and AR playback of the emerging MPEG video-based point cloud compression standard," *Nokia Technologies; IBC: Helsinki, Finland*, 2019.

[41] "Google. DRACO : 3D DATA COMPRESSION ." (), [Online]. Available: `https://google.github.io/draco/`.

[42] M. Bui, L.-C. Chang, H. Liu, Q. Zhao, and G. Chen, "Comparative study of 3D point cloud compression methods," in *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, 2021, pp. 5859–5861.

[43] T. Huang and Y. Liu, "3D point cloud geometry compression on deep learning," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 890–898.

[44] M. Quach, G. Valenzise, and F. Dufaux, "Improved deep point cloud geometry compression," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2020, pp. 1–6.

[45] J. Wang, H. Zhu, H. Liu, and Z. Ma, "Lossy point cloud geometry compression via end-to-end learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4909–4923, 2021.

[46] I.-T. J.1631, "Functional requirements of E2E network platforms to enhance the delivery of cloud-VR services over integrated broadband cable networks," 2021.

[47] T. L. Burnett, "61-1: Invited paper: Light-field display architecture and the challenge of synthetic light-field radiance image rendering," in *SID Symposium Digest of Technical Papers*, Wiley Online Library, vol. 48, 2017, pp. 899–902.

[48] D. Xu, A. Zhou, X. Zhang, G. Wang, X. Liu, C. An, Y. Shi, L. Liu, and H. Ma, "Understanding operational 5G: A first measurement study on its coverage, performance and energy consumption," in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, 2020, pp. 479–494.

[49] J. Rischke, P. Sossalla, S. Itting, F. H. Fitzek, and M. Reisslein, "5G campus networks: A first measurement study," *IEEE Access*, vol. 9, pp. 121 786–121 803, 2021.

[50] A. Narayanan, E. Ramadan, R. Mehta, X. Hu, Q. Liu, R. A. Fezeu, U. K. Dayalan, S. Verma, P. Ji, T. Li, *et al.*, "Lumos5G: Mapping and predicting commercial mmWave 5G throughput," in *Proceedings of the ACM Internet Measurement Conference*, 2020, pp. 176–193.

[51] I. F. Akyildiz, C. Han, Z. Hu, S. Nie, and J. M. Jornet, "Terahertz band communication: An old problem revisited and research directions for the next decade," *IEEE Transactions on Communications*, vol. 70, no. 6, pp. 4250–4285, 2022.

[52] C. Liaskos, L. Mamatas, A. Pourdamghani, A. Tsioliaridou, S. Ioannidis, A. Pitsillides, S. Schmid, and I. F. Akyildiz, "Software-defined reconfigurable intelligent surfaces: From theory to end-to-end implementation," *Proceedings of the IEEE*, vol. 110, no. 9, pp. 1466–1493, 2022.

[53] X. Liu, Y. Deng, C. Han, and M. Di Renzo, "Learning-based prediction, rendering and transmission for interactive virtual reality in RIS-assisted terahertz networks," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 2, pp. 710–724, 2021.

[54] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[55] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.

[56] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for vqa," *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 553–557, 2021.

[57] Y. Bo, Y. Duan, S. Shao, and M. Tao, "Learning Based Joint Coding-Modulation for Digital Semantic Communication Systems," in *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2022, pp. 1–6.

[58] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[59] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva, "Imaginator: Conditional spatio-temporal GAN for video generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1160–1169.

[60] D. He, C. Westphal, and J. Garcia-Luna-Aceves, "Joint rate and FOV adaptation in immersive video streaming," in *Proceedings of the 2018 Morning Workshop on Virtual Reality and Augmented Reality Network*, 2018, pp. 27–32.

[61] L. Wang, Z. Cao, Z. Cui, C. Cao, and Y. Pi, "Negative latency recognition method for fine-grained gestures based on terahertz radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7955–7968, 2020.

[62] X. Fang, W. Feng, Y. Chen, N. Ge, and Y. Zhang, "Joint Communication and Sensing toward 6G: Models and Potential of Using MIMO," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4093–4116, 2023.

[63] R. Li, K. Makhijani, and L. Dong, "New IP: A data packet framework to evolve the internet," in *2020 IEEE 21st International Conference on High Performance Switching and Routing (HPSR)*, IEEE, 2020, pp. 1–8.

[64] J. Garcia-Luna-Aceves, D. He, and C. Westphal, "Network support for ar/vr and immersive video application: A survey," in *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, ICETE 2018*, vol. 1, 2018.

[65] Y. Xu, F. Dang, R. Xu, X. Chen, and Y. Liu, "Lsync: A universal event-synchronizing solution for live streaming," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, IEEE, 2022, pp. 2188–2197.

[66] Z. Yuan, T. Bi, G.-M. Muntean, and G. Ghinea, "Perceived synchronization of mulsemedia services," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 957–966, 2015.

[67] O. A. Ademoye and G. Ghinea, "Synchronization of olfaction-enhanced multimedia," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 561–565, 2009.

[68] I. F. Akyildiz and W. Yen, "Multimedia group synchronization protocols for integrated services networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 162–173, 1996.

[69] P. V. Rangan, H. M. Vin, and S. Ramanathan, "Communication architectures and algorithms for media mixing in multimedia conferences," *IEEE/ACM Transactions On Networking*, vol. 1, no. 1, pp. 20–30, 1993.

[70] R. Steinmetz, "Synchronization properties in multimedia systems," *IEEE Journal on selected areas in communications*, vol. 8, no. 3, pp. 401–412, 1990.

[71] T. D. C. Little and A. Ghafoor, "Multimedia synchronization protocols for broadband integrated services," *IEEE Journal on selected Areas in Communications*, vol. 9, no. 9, pp. 1368–1382, 1991.

[72] Y. Ishibashi and S. Tasaka, "A group synchronization mechanism for live media in multicast communications," in *GLOBECOM 97. IEEE Global Telecommunications Conference. Conference Record*, IEEE, vol. 2, 1997, pp. 746–752.

[73] H. Fu, M. Sun, B. He, J. Li, and X. Zhu, "A survey of traffic shaping technology in internet of things," *IEEE Access*, vol. 11, pp. 3794–3809, 2023.

[74] D. Xu, T. Li, Y. Li, X. Su, S. Tarkoma, T. Jiang, J. Crowcroft, and P. Hui, "Edge intelligence: Empowering intelligence to the edge of network," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1778–1837, 2021.

[75] J. Blocklove, S. Garg, R. Karri, and H. Pearce, "Chip-Chat: Challenges and Opportunities in Conversational Hardware Design," *arXiv preprint arXiv:2305.13243*, 2023.

[76] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large Language Models for Telecom: The Next Big Thing?" *arXiv preprint arXiv:2306.10249*, 2023.

[77] D. Rossi and L. Zhang, "Landing AI on Networks: An equipment vendor viewpoint on Autonomous Driving Networks," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 3670–3684, 2022.

[78] C. Benzaid and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions," *Ieee Network*, vol. 34, no. 2, pp. 186–194, 2020.

[79] A. C. da Silveira and C. A. S. Santos, "Ongoing Challenges of Evaluating Mulsemedia QoE," in *Proceedings of the 2nd Workshop on Multisensory Experiences-SensoryX'22*, SBC, 2022.

[80] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can terahertz provide high-rate reliable low-latency communications for wireless VR?" *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9712–9729, 2022.

AUTHORS

Ian F. Akyildiz received BS, MS, and PhD degrees in electrical and computer engineering from the University of Erlangen–Nurnberg, Germany, in 1978, 1981, and 1984, respectively. Since 1989, he has been the founder and President of Truva Inc., a consulting company based in Georgia, USA. He has also been a member of the Advisory Board at the Technology Innovation Institute (TII) Abu Dhabi, United Arab Emirates, since June 2020. He has been the founder and Editor-in-Chief of the newly established of International Telecommunication Union Journal on Future and Evolving Technologies (ITU J-FET) since August 2020.

He served as the Ken Byers Chair Professor in Telecommunications, the Past Chair of the Telecom Group at the ECE, and the Director of the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, from 1985 to 2020. He had many international affiliations during his career and established research centers in Spain, South Africa, Finland, Saudi Arabia, Germany, Russia, India, and Cyprus. Dr. Akyildiz is an IEEE Life Fellow and an ACM Fellow. He has received numerous awards from IEEE, ACM, and other professional organizations, including the Humboldt Award from Germany. In August 2022, according to Google Scholar his H-index is 133 and the total number of citations to his articles is more than 136K+. His current research interests include 6G/7G wireless systems, terahertz communication, reconfigurable intelligent surfaces, nanonetworks, Internet of Space Things/CUBESATs, Internet of Bio-Nano Things, molecular communication, and underwater communication.
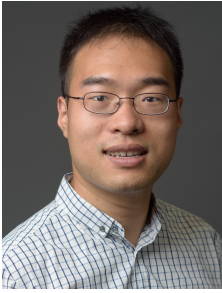
Hongzhi Guo received his Ph.D. degree from the University at Buffalo in 2017. Currently, he is an assistant professor in the School of Computing at the University of Nebraska-Lincoln. His broad research agenda is to develop foundations for wireless networks and networked robotics to automate dangerous dirty dull tasks in extreme environments. He received the NSF CAREER award in 2022.
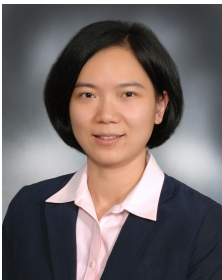
Rui (April) Dai is an associate professor in the Department of Electrical and Computer Engineering at the University of Cincinnati, Ohio, USA. She received her BS in electronics and information engineering and MS in communications and information systems from Huazhong University of Science and Technology, Wuhan, China, in 2004 and 2007, respectively. She received her PhD degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA in 2011. She was a postdoctoral fellow at the Center for Assistive Technology and Environmental Access of Georgia Tech from 2011 to 2012. She was an assistant professor at the Department of Computer Science at North Dakota State University from 2012 to 2014 and an assistant professor in the Department of Electrical Engineering and Computer Science at University of Cincinnati from 2014 to 2020. Her current research interests include multimedia communications and networking, wireless sensor networks, and cyber-physical systems.

Wolfgang Gerstacker received the Dipl.-Ing. degree in electrical engineering and the Dr.-Ing. and Habilitation degrees from Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Erlangen, Germany, in 1991, 1998, and 2004, respectively. Since 2002, he has been with the Chair of Mobile Communications and Institute for Digital Communications, FAU Erlangen-Nürnberg, where he is currently a professor. He has conducted various projects with partners from industry. His research interests include the broad areas of digital communications and statistical signal processing, THz communications, 5G and beyond, machine learning for wireless communications and wireless sensor networks. He is a member of the Executive Editorial Committee of IEEE Transactions on Wireless Communications. He was a recipient of several awards, including the Research Award of the German Society for Information Technology (ITG) in 2001, the EEEfCOM Innovation Award in 2003, the Vodafone Innovation Award in 2004. He is an editor of Computer Networks (Elsevier). He has served as a member of the Editorial Board for IEEE Transactions on Wireless Communications, Physical Communication (Elsevier), and EURASIP Journal on Wireless Communications and Networking, and as a guest editor for several IEEE journals and magazines.