INVITED PAPER

The Road to Immersive Communication

This comprehensive description of rich-media communications defines how quality of experience should be understood. The increasing role of new modalities such as augmented reality and haptics is discussed.

By John G. Apostolopoulos, *Fellow IEEE*, Philip A. Chou, *Fellow IEEE*, Bruce Culbertson, *Member IEEE*, Ton Kalker, *Fellow IEEE*, Mitchell D. Trott, *Fellow IEEE*, and Susie Wee, *Fellow IEEE*

ABSTRACT | Communication has seen enormous advances over the past 100 years including radio, television, mobile phones, video conferencing, and Internet-based voice and video calling. Still, remote communication remains less natural and more fatiguing than face-to-face. The vision of immersive communication is to enable natural experiences and interactions with remote people and environments in ways that suspend disbelief in being there. This paper briefly describes the current state-of-the-art of immersive communication, provides a vision of the future and the associated benefits, and considers the technical challenges in achieving that vision. The attributes of immersive communication are described, together with the frontiers of video and audio for achieving them. We emphasize that the success of these systems must be judged by their impact on the people who use them. Recent high-quality video conferencing systems are beginning to deliver a natural experience—when all participants are in custom-designed studios. Ongoing research aims to extend the experience to a broader range of environments. Augmented reality has the potential to make remote communication even better than being physically present. Future natural and effective immersive experiences will be created by drawing upon intertwined research areas including multimedia signal processing, computer vision, graphics, networking, sensors, displays and sound reproduction systems, haptics, and perceptual modeling and psychophysics.

Manuscript received May 22, 2011; revised September 19, 2011; accepted October 20, 2011. Date of publication February 16, 2012; date of current version March 21, 2012.
J. G. Apostolopoulos, **B. Culbertson**, and **M. D. Trott** are with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (e-mail: john_apostolopoulos@hp.com).
P. A. Chou is with Microsoft Research, Redmond, WA 98052 USA.
T. Kalker is with Huawei Innovation Center, Santa Clara, CA 95050 USA.
S. Wee is with Cisco Systems, San Jose, CA 95134 USA.

Digital Object Identifier: 10.1109/JPROC.2011.2182069

KEYWORDS | Immersive environments; signal processing; telepresence; video conferencing

I. INTRODUCTION

Communication technology strives to pull us together faster than transportation technology and globalization push us apart. Today we can communicate with almost anyone else in the world by voice, e-mail, text, and even video. This has permitted us to serve our customers far from where we work, work far from where we live, and live far from our friends and families [1]. Yet today's communication experiences are impoverished compared to face-to-face interaction. Thus, while advances in communication have increased the quantity of connections we can maintain, they have failed to preserve their quality.

Humans are social animals, and have evolved to interact with each other most effectively face-to-face. As humans, we continue to meet in person whenever feasible. We continue to meet in conference rooms; we continue to commute to work; we continue to fly to conferences, trade shows, and weddings [2]. The reasons are clear. Colocation enables us to exhibit and interpret various nonverbal signals and cues, such as touch, proximity, position, direction of gesture, eye contact, level of attention, etc. [3]–[5]. It enables us to greet each other with a handshake or embrace, dine together, and appreciate objects in a shared environment. Today's technologies do not permit us to perform many of these social interactions.

A critical point is that the quality of an immersive communication system is judged by its impact on the humans who use it. Unfortunately, it is exceedingly difficult to develop objective metrics to assess that impact.

Therefore, designing, analyzing, and predicting the future of immersive communication systems is as much art as science.

Our need for more effective ways to communicate with remote people has never been greater. Numerous societal conditions are aligning to create this need: the urgency of reducing environmental impact, the demand to reduce travel costs and fatigue, the need for richer collaboration in an ever more complex business environment, and the difficulty of travel during natural disasters such as volcanic eruptions and influenza outbreaks.

In this paper, we attempt to indicate how, despite the wide gap between today's communication technologies and the experience of actually being in another location, this gap will eventually be substantially reduced. Technology advances faster than biology, and soon it will be practical to deliver more bits per second to an observer than can pass through the sensory cutset separating the environment from the nervous system. The human visual system can absorb only about 10 Mb/s; the haptic system about 1 Mb/s; auditory and olfactory systems about 100 kb/s each; and gustatory system about 1 kb/s [6]. In contrast, fiber-to-the-home is already delivering on the order of 10 Mb/s, while links in the Internet backbone are approaching 1 Tb/s [7]. With such vast information rates already possible, the question is not how to communicate information; rather, it is about transduction: how to capture and render that information in ways that match the human system, and more fundamentally, what we want our communication and remote experiences to be like.

The answer to the latter question, to the extent that popular culture reflects our desires, is *immersive*. For the purposes of this paper, *immersive communication* means exchanging natural social signals with remote people, as in face-to-face meetings, and/or experiencing remote locations (and having the people there experience you) in ways that suspend disbelief in *being there*. There is transparency, fluidity of interaction, a sense of presence. *Star Trek*'s Holodeck, *Star Wars*' Jedi council meetings, *The Matrix*'s matrix, and *Avatar*'s Pandora all illustrate visions of immersive communication.

A key feature of immersive communication is the ability for participants to interact with the remote environment, and to detect and evaluate that interaction via their senses. Visually, as we move, we see the world from different perspectives, the world reacts, and people we are looking at react, helping us feel as if we are part of the scene. Similarly, we hear the scene as we turn around, the scene adapts to us, and we hear the results. This two-way interaction gives us a feeling of immersion. Other senses are analogous, but are technically harder to convey. Our vestibular sense (balance and acceleration) is coupled deeply to our visual system, and inconsistency between these two can be profoundly distressing. Touch is the next most important sense after vision and hearing, but we will only briefly discuss it in this paper. Communication of smell and taste is in its infancy, and we will not dwell upon it here.

What humans are able to sense at a given moment is a tiny slice of the full environment. For example, high-resolution detail cannot be perceived outside the current center of gaze, loud sounds mask quiet ones, and polarized light cannot be directly sensed at all. Exploiting these features of human perception can vastly simplify immersive communication.

Immersion evokes a state of mind or emotion of being in another place, or of others being in your place; that is, suspending your disbelief in being elsewhere, as when you are immersed in a book, movie, or video game. These examples indicate the power of the human ability to "fill in" missing or inconsistent audio, video, and related cues, a power exploited by present and likely future immersive systems.

We do not posit a single ideal notion or degree of immersion. Far from it: there will always be a range of communication needs and purposes, each served best by different means. Even simple text messaging is likely to survive for good reasons. Certain immersive forms of communication may be undesirable due to lack of privacy and attention requirements [8], [15]. Price, portability, power, and other constraints will prevail. For example, a person may shift between a handheld, a wall-sized, and a head-mounted display, each providing a rather different degree and style of immersion. Virtual spaces need not depict a perfectly consistent physical geometry; indeed, when combining several remote sites, nonphysical layouts may be best. Thus diverse and partial illusions of immersion will continue to be achieved by ingenious blending of immutable elements of the physical environment with the virtual.

Similarly, there is no single technical roadmap to immersive communication. The invention, for example, of ideal light field cameras and displays would not suddenly enable immersive communication for a smartphone user; the camera will not be positioned well, and even a near-perfect 9-cm screen remains but a small window into another space. To enable immersive experiences in this and many other scenarios, technical innovations must be paired with equally challenging nontechnical ones in areas such as creative design.

In addition to natural ("as good as being there") communication enabled by immersion, supernatural ("better than being there") communication will also be important. For example, it will be possible to look or listen across large distances, rewind or accelerate time, speed through space, record experiences, or provide translation abilities. A politician could make eye contact with everyone, simultaneously. An orchestra, limited today by acoustic delays to perhaps a hundred people, could swell to thousands of networked participants.

Many open questions remain about how to foster the illusion of immersion. Transparency and fluidity of interaction likely plays an important role. For example, joining

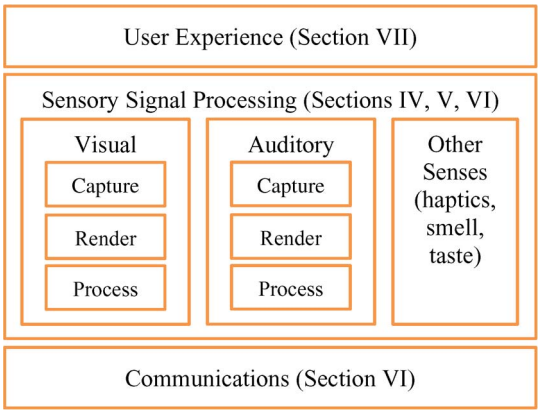


Fig. 1. ***The elements of an immersive communication system.***

an interaction today is far different from physically entering a room or encountering someone at a party. Immersion could also be interpreted to include being bathed in a sea of data, such as text message feeds from compatriots during a revolution. Other extrasensory aspects may also be important. Nevertheless, in this paper, we will focus primarily on aspects of immersive communication involving perception and the senses, particularly auditory and visual information.

We believe that highly immersive communication will become ubiquitous over the coming decades, as the cost of computation, bandwidth, and device resolution continues to drop. The goal of this paper is to illustrate possible technical paths to this end.

An immersive communication system consists of the user experience, sensory signal processing, and communications as shown in Fig. 1. The sense of immersion that a user experiences is driven by the quality of experience (QoE) that the system provides. The QoE rests, in turn, on how video, audio, and other sensory signals are captured, processed, and rendered. Ultimately, in an immersive environment, these signals rely for their reproduction on more bandwidth and more sophisticated communication than is common today.

The outline of the paper is as follows. After a brief history of immersive communication in Section II, we lay out a framework for thinking about various categories of immersive scenarios and applications in Section III, including a discussion on augmented reality, mobility, robots, and virtual worlds. We establish the technical core of the paper in Sections IV–VI, which lay out the fundamental notions of sensory signal processing and communications, specifically in the immersive rendering, acquisition, and compression of visual and auditory immersive communication. Special emphasis is paid to the interaction between rendering and acquisition due to interactivity requirements. We also briefly discuss in Section VI the important topic of haptics and its compression and communication. Section VII discusses the role of human perception on QoE, including peripheral awareness, consistency between local and remote spaces, and 3-D cues. Section VIII provides concluding remarks.

Two types of scenarios are helpful to consider throughout, as they illustrate a wide range of human interaction and motivate corresponding forms of immersive communication. These scenarios include small group collaborations such as business meetings and large social gatherings such as weddings. Business meetings typically have, say, 3–30 participants in a relatively constrained environment with a relatively limited set of interactions. Weddings are usually an order of magnitude bigger, and present a much wider range of social interactions. Immersive communication for business meetings is on the immediate horizon, while immersive communication for weddings may be more than a generation away from full technical feasibility and social acceptance. However both scenarios, as well as other scenarios of interest (such as lectures, dining, sporting events, gaming, conferences, trade shows, or poster sessions) share a common need for accurate communication of social signals and interactions with the environment, made possible by immersive communication.

II. BRIEF HISTORY OF IMMERSIVE COMMUNICATION

Long before it could be realized, people dreamed of technology that would communicate sight as well as sound. The invention of the telephone spawned a flurry of examples, including George du Maurier's 1878 cartoon in Punch Almanack [9], shown in Fig. 2. Fanciful depictions in movies and on TV, including *Metropolis* (1927), *Modern Times* (1936), *2001: A Space Odyssey* (1968), *Star Trek* (1966), and *The Jetsons* (1962), offer hints of people's hopes and expectations for immersive communication, and

Fig. 2. ***George du Maurier anticipates immersive communication in this 1878 cartoon.***

actually inspire developers of the technology. Scientists have speculated more seriously on the possibilities of such technology. Alexander Graham Bell [10] wrote about seeing with electricity, Ivan Sutherland [11] expounded on ideal displays and the interaction they would enable, and Marvin Minsky [8] refined and popularized the vision of *telepresence*. Minsky's telepresence focused on high-quality sensory feedback of sight, sound, and touch/manipulation, which would enable people to work remotely, for example, perform an operation or work in a damaged nuclear reactor.

The telephone and television are critical enabling technologies for immersive communication. A number of inventors contributed to the telephone, with Alexander Graham Bell receiving the key patent in 1876. Many television-like devices were developed in the early 20th century. A system demonstrated in 1928 [12] used electronic scanning for both image capture and display, anticipating virtually all succeeding television systems. One of the first television broadcasts was the 1936 Berlin Olympic Games.

Reasonable people will differ on when the history of actual immersive communication systems begins, due to varying judgments of the amount of immersiveness a system needs to qualify. The German Reichpost, or post office, built one of the first public video conference networks, connecting Berlin, Nuremberg, Munich, and Hamburg, operating from 1936 to 1940. AT&T demonstrated its videophone, trademarked *Picturephone*, at the 1964 New York World's Fair. The product was dropped after a few years due to its low video quality and high (US$16 per three minutes) cost. Though quality and cost gradually improved, they continued to limit the commercial success of these systems into the 1990s. At the end of that decade, systems with high-quality, life-size video began to appear from companies like TeleSuite, Teliris, Hewlett-Packard, Polycom, and Cisco, albeit still at a high price. Microsoft NetMeeting, Apple iChat, and Skype, which added video to PC conferencing beginning in 1995, 2003, and 2006, respectively, provided three additional features whose absence had inhibited widespread use of many earlier systems: ubiquity, ease of use, and extremely low cost. Universal Mobile Telecommunications System (UMTS) enabled video calling on cell phones in the early 2000s. International standards have been created by the ITU, ISO/IEC, IETF and other standardization bodies to foster interoperability across endpoints developed by different manufacturers and networks operated by different providers.

With new applications appearing at an accelerating pace, immersive communication is poised to advance explosively, aided by progress in computation, bandwidth, and device capabilities. It is also significantly benefiting from the flexibility of the generic data carried by IP packets, in sharp contrast to analog television signaling. Though initially often used by communication system developers to mimic earlier video functionality, IP is enabling a host of enhanced and entirely new modes of communication. Cave Automatic Virtual Environments (CAVEs) [13] surround users with displays on the walls, and sometimes even the ceiling and floor, literally immersing them in imagery. Haptic systems convey the sense of touch, for example, enhancing computer games and helping pilots fly airplanes remotely. The concept of virtual reality, which ultimately would stimulate the senses to create an experience indistinguishable from reality, significantly predates actual immersive communication systems. Augmented reality enhances a depiction of a real environment, for example, annotating live video of a mechanism to facilitate a repair.

III. CLASSES OF IMMERSIVE SYSTEMS

This section points out various classes of immersive systems, by means of the framework illustrated in Fig. 3, which shows a spectrum of communication and collaboration systems ranging from communication systems designed to support *natural conversation* between people, e.g., using voice and video, to collaboration systems designed to support *sharing information* between people. The framework shows how currently existing systems anywhere along this spectrum have become and are continuing to become increasingly immersive.

Specifically, communication systems for natural conversation have progressed from phone calls, to multiparty video conferences, to high-end telepresence systems such as HP Halo and Cisco Telepresence that share voices and life-size video at a quality that gives participants the feeling of being in the same room. This progression increases the presence and connection between remote participants and

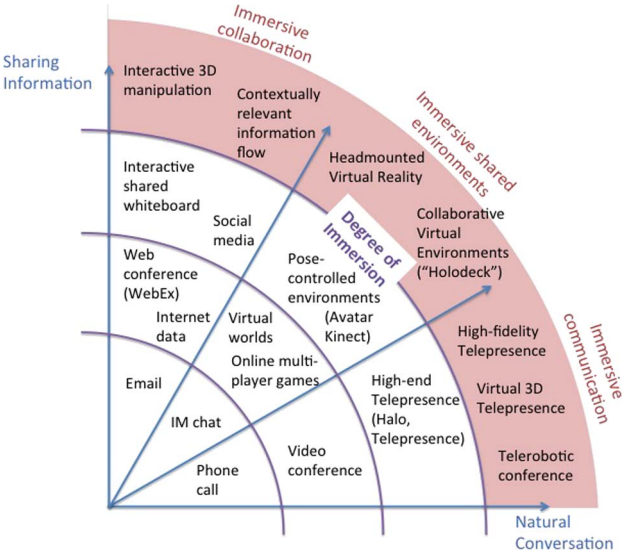


Fig. 3. ***Communication and collaboration framework.***

leads to more advanced *immersive communication* systems, which are evolving in a number of directions. One form is a high-fidelity telepresence system in which space is virtually extended by adjoining remote locations through ideal, wall-sized displays. Another form of immersive communication gives remote participants a virtual 3-D presence by capturing and analyzing their movements and rendering them locally using a type of augmented reality. A third form gives remote participants a physical presence through telerobotic conferences. In these systems, the video, voice, and movements of a remote participant are captured and analyzed, and are then emulated by a physical robot. The remote participant may even control the robot's movement to navigate through an area [16].

Collaboration systems for sharing information have progressed from telegraphs and e-mail to web conferencing systems that allow the sharing of documents, applications, and computer desktops to interactive shared whiteboards. Another degree of information sharing is achieved by augmenting the collaboration topic at hand with digital data mined from the Internet and social media tools such as Facebook and Twitter, which allow large numbers of participants to contribute to a conversation both synchronously and asynchronously. We can imagine these tools evolving into truly immersive systems in which augmented reality techniques are used to share data between participants, for example, allowing interactive 3-D manipulation of objects that is augmented by an inflow of contextually relevant digital information.

One can consider a continuum of systems that combine natural conversation and sharing information to create shared environments for collaboration. IM chat can be considered one such system, as a chat room is a shared space where people can go to share information. Further immersion is achieved in allowing participants to join a shared environment in online multiplayer games such as Xbox Live and in virtual worlds such as Second Life. Pose controlled environments such as Avatar Kinect allow people to be represented by avatars that emulate their expressions and movements. Advanced *immersive and shared environments* can take different forms: head-mounted virtual reality systems allow a participant to more fully experience the shared environment. An ultimate goal is to mimic collaborative virtual environments depicted in science fiction such as the Holodeck in *Star Trek*.

Thus, immersive communication is evolving in a variety of directions. One possible goal is to emulate reality, in which case high fidelity, realistic renderings are important and spatial relationships must be maintained. Another possible goal is to achieve a state of immersive communication that goes beyond reality, for example, enabling millions of people to attend a lecture or a music concert in which everyone has a front row seat.

The next few sections address technologies expected to underlie a large variety of such immersive communication systems.

IV. VISUAL RENDERING, CAPTURE, AND INTERACTION

Visual information is a key element of immersive communication. This section begins by considering the display and rendering of visual information, and the desired attributes to provide an immersive experience. Given these attributes we examine the capture of the visual information. The desire to provide natural interaction between people leads to additional constraints on the capture and rendering.

A. Visual Rendering

The goal of the visual rendering is to provide an immersive visual experience for each viewer. An ideal display places a light field across the pupil of each eye as if the viewer were viewing the scene directly. This has several key attributes: each eye of each viewer sees an appropriate view (stereoscopy), the view changes as the viewer moves (motion parallax), and objects move in and out of focus as the viewer changes his or her focal plane. Peripheral vision is also very important; an ideal display fills the field of view. If the display is very close to a viewer, for example, head-mounted, it can completely fill the field of view in a compact space. More distant displays must be large enough to wrap around the viewer to achieve such an effect. Ideal displays can thus be either large or small, with the caveat that the larger, more distant displays can be occluded by objects in the real world and hence such displays may not always be able to strictly control the light reaching the viewer.

A *light field* [23], [24], or more generally, a *plenoptic function* [17], expresses the intensity of light rays within a space. Formally, the 7-D plenoptic function $P(x, y, z, \theta, \phi, \lambda, t)$ is the intensity of the light ray passing through spatial location (x, y, z), at angle (θ, ϕ), for wavelength λ, and at time t. Polarization and phase (important for holography) can also be included by considering the vector-valued function $\vec{P}(x, y, z, \theta, \phi, \lambda, t)$, where $\vec{P}$ is the vector strength of the electric (or magnetic) field in the plane perpendicular to the direction of the light ray.

Since light travels through free space in a straight line (ignoring relativistic effects), essentially all light rays that pass through a volume of space also intersect a nearby plane. Hence, the plenoptic function within the volume can be characterized by the value of plenoptic function on the plane. Parameterizing positions on the plane by x and y, and repressing λ and t for simplicity, we get the simplified 4-D plenoptic function $P(x, y, \theta, \phi)$, known as the *light field* or *lumigraph* in computer vision and computer graphics [23], [24].

An ideal flat display, whether large or small, is able to reproduce a light field $P(x, y, \theta, \phi)$ across its surface, such as would be produced across an open window. That is, each spatial location (x, y) on the surface of the ideal display is able to cast different light rays in different directions (θ, ϕ), as illustrated in Fig. 4 (left).

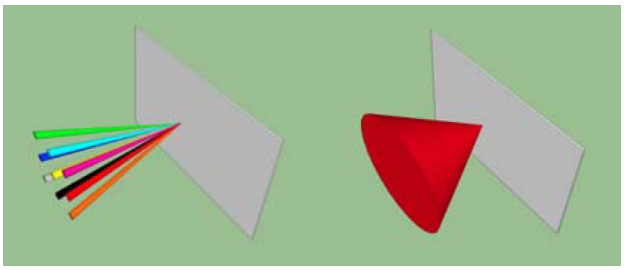


Fig. 4. *For each spatial location in a rectangular display, an ideal light field display (left) emanates different light rays in different directions, as compared to a conventional display (right) which emanates the same color in all directions.*

Conventional displays, in contrast to the aforementioned ideal display, correspond to a 2-D matrix of light emitters, referred to as pixels, where each pixel emits a single color in *all* directions (θ, ϕ) corresponding to a large cone of light, as illustrated in Fig. 4 (right).

A stereoscopic 3-D display improves on a conventional display by providing two views for each (x, y) location: one view for the right eye and a second for the left eye. This is often accomplished using temporal, polarization, or wavelength multiplexing of the emitted light with eyeglasses that demultiplex the light so the appropriate view reaches each eye [32]. Multiple viewers with glasses will see the same images, whether or not the images are appropriate for their relative position. A multiview display extends two-view stereo display to N views, where N is typically on the order of tens of views. In an N-view multiview display each (x, y) location emits N light cones in N directions. Autostereoscopic displays provide two or more views without requiring glasses. This is often accomplished by placing a parallax barrier or lenticular array in front of a conventional 2-D matrix of pixels, which directs light from different pixels in different directions [32]. Note that if the 2-D matrix has a total of M pixels, and there are N views, each view has M/N pixels. Therefore, achieving a large number of views and a high resolution for each view requires a large number of pixels.

A stereoscopic two-view display forces the viewer to focus on the surface of the screen even when viewing an object depicted at a different depth. Indeed, it is the display, not the viewer, which selects the parts of the scene that are in focus. In contrast, a light field display allows the viewer to adapt focus and track depth in a natural manner, as if the viewer were looking through an open window.

The limitations of human perception may allow the light field to be heavily sampled across each dimension. For example, it may be sufficient to sample the angular dimensions (θ, ϕ) to resolve different rays at 1/2 pupil width. Assuming a 2-mm pupil and a display distance of 1 m, covering all 180° of viewing angle amounts to 3000 rays horizontally and 3000 vertically, or 9 000 000 rays per (x, y) coordinate. Such a display bathes a room in light, further enhancing the illusion of a window. The complexity can be dramatically reduced by directing light only to the pupils of the viewers, for example, by using a camera to track the viewers and their pupils, thereby reducing the required number of rays by many orders of magnitude.

The feeling of immersion can be improved by providing a large field of view, for example, by employing large displays. Current display fabrication technologies, such as for LCD or plasma displays, are not amenable to large sizes, as the yield decreases with increasing size, leading to higher cost. Therefore, large displays are currently produced using projector systems, sometimes involving multiple projectors, or a mosaic of smaller LCD displays. A promising approach for future economical manufacture of large area displays is through the use of roll-to-roll (R2R) manufacturing techniques where the active matrix electronics backplane for displays is imprinted on a roll of plastic substrate. A major challenge for R2R techniques has been preserving alignment during the various fabrication steps, which induce stretching, shearing, etc., across the large, flexible, plastic substrate. The development of self-aligned imprint lithography (SAIL) overcomes this problem and provides submicron alignment [18], [19]. Displays made on plastic substrate also facilitate the creation of flexible displays or displays with new form factors, e.g., large, curved displays that surround the viewers, further increasing their sense of immersion.

Volumetric displays [20] provide a benefit over a planar display in that the viewers can walk around the display. Flexible displays may provide the ability to produce cylindrical displays, providing a similar effect to volumetric displays.

A head-mounted display can completely fill the field of view, and can replace or augment the ambient light, for example, to insert a virtual person into an otherwise empty chair. Precise, low-latency head-tracking coupled to the rendering system is an essential component for such a display. As head-mounted displays become more comfortable, higher performing, and less visually intrusive to others, they will likely occupy larger and larger niches in immersive communication.

B. Capture

A conceptually ideal visual capture system, complementary to an ideal display that emulates a large window, captures all of the visual information that would go through this window, namely the light field or 4-D plenoptic function $P(x, y, \theta, \phi)$ (again repressing color and time for simplicity).

It is useful to compare this idealized capture device with a conventional camera. A conventional camera uses a lens to focus the visual information on a 2-D sensor, corresponding to a 2-D matrix of sensing elements. Each sensor element captures the integral of the light impinging on it, irrespective of direction. Therefore, a conventional camera

captures only a 2-D projection of the 4-D plenoptic function. A stereo camera incorporates two lenses and two 2-D sensors, to capture two projections of the 4-D plenoptic function. A camera array can capture multiple views of the scene.

Efforts to capture the plenoptic function include placing a 2-D lens array in front of a 2-D sensor. The lens array provides spatial sampling of (x, y), and each lens then focuses different angles onto different sensors in the 2-D sensor, thereby capturing both spatial and angular information. Note that the product of spatial and angular sampling is bounded above by the number of sensors in the 2-D sensor. As the lens array gets larger, and the individual lenses get smaller, and the 2-D sensor increases in resolution, this approaches a light-field capture device, and is sometimes referred to as a plenoptic camera [21], [22].

The above discussion highlights the need to understand the minimum rate required to sample the plenoptic function, analogous to the Nyquist sampling theory for 1-D signals. A number of studies have examined the plenoptic function in the spectral domain under different assumptions, and the tradeoffs between camera sampling, depth of the scene, etc., [25]–[27]. For example, Do *et al.* [28] showed that for the simplistic case of painting a bandlimited image on a flat surface in 3-D space, the plenoptic function is bandlimited. However, they also showed that the plenoptic function is not bandlimited unless the surface is flat. They provide rules to estimate the bandwidth depending on the maximum and minimum scene depths, maximum slope of the surface, and maximum frequency of the painted signal.

The future will see increased use of capture devices that provide *more* information than what a human can see through a window. For example, capturing other spectrum bands such as infrared or ultraviolet may be useful in some applications including enhancement. Also, depth cameras extend the conventional camera or the light field camera to include an associated depth per pixel. Depths can be estimated using a variety of techniques including passive stereo depth estimation, and active time-of-flight or structured light patterns [29], [31] where the scene is illuminated with typically a near-infrared signal. Both passive and active techniques have strengths and weaknesses, which are somewhat complimentary, and their strengths can be combined by fusing both passive and active depth estimation [30]. Accurate depth information simplifies a variety of applications including foreground/background segmentation, pose estimation, gesture recognition, etc. [31].

C. Two-Way Interaction

Two-way, and generally N-way, interaction between people leads to additional requirements on capture and rendering. For example, face-to-face discussions employ important nonverbal signals such as eye contact, gaze direction, and gestures such as pointing. However, conventional video conferencing systems have the camera above the display, which does not provide good eye contact or gaze awareness. If an individual looks at the display the camera will capture a view in which the individual appears to be looking down. Looking at the camera corrects the remote view but then the local viewer cannot concentrate on the display. The directions of gaze and gestures are also incorrectly expressed using current systems. This problem occurs because the appropriate position of the camera is behind the display, looking through it. Making a hole in a display and placing a conventional camera at the hole does not solve the problem as the viewpoint is too close. These problems have motivated the design of systems with see-through displays to provide correct eye contact and gaze awareness [33], [50].

In scenarios such as wedding receptions where participants are free to move about, the desired viewpoint for remote participants is not known *a priori*, hence camera placement becomes problematic. In such cases, view synthesis algorithms may be required to estimate the desired views. Such techniques remain an active and highly challenging field of research.

D. Summary

The future promises many advances that will improve the visual immersive experience. Visual rendering will move from pixels, which emit the same color in all directions to light field displays, which emit different colors in different directions. This would enable glasses-free 3-D viewing, while overcoming the vergence–accommodation conflict that afflicts conventional stereoscopic 3-D displays. Capture systems will support the capture of these light fields. In addition, see-through display systems will provide correct eye contact and gaze awareness. It is important to emphasize that these systems involve an immense amount of data to capture, process, compress, transport, decode, and display. The implications on compression and transport are discussed in Section VI. The challenge of working with the immense amount of data associated with entire light fields motivates techniques to intelligently select a subset of information to capture or render. For example, if the goal is to reproduce the light field at the location of the human eyes, then only those associated light rays need to be captured. In addition, sometimes the light rays needed for display may not have been captured, leading to the necessity of interpolating the missing rays from those available, which relates to the well-known problem of view interpolation [34], [35]. It is also noteworthy that future displays may include both emitters and sensors on the same surface, where the surface can simultaneously act as a display and a capture device [36].

V. AUDITORY RENDERING, CAPTURE, AND INTERACTION

Two-way audio connectivity is universally compelling, as witnessed by the remarkable uptake first of wired

telephony and then cell phones. Paradoxically, however, once basic connectivity is available, humans seem remarkably tolerant of bad audio quality; historic wired telephony is passable for voice but it can hardly be considered high fidelity, while cell phones are noticeably worse. It remains commonplace, for example, to endure wired and wireless phone calls where the remote party is marginally intelligible. A range of human experiences need better audio, with fidelity beyond the distortions of today's phone calls, and going beyond single-channel audio to some form of multichannel audio that provides a sense of spatial richness and immersion not fully achievable even with today's research systems. While the beneficial effects are difficult to quantify, it appears that immersive audio grows in importance when the number of participants increases, when the interaction entails moving through a real or virtual space, or when emotion plays a role, for example, when teaching or persuading.

In a formal meeting, immersive audio allows participants to quickly determine who is talking, to detect the subtle signals that direct conversation flow (e.g., throat clearing), and to discern mood (e.g., a snort of derision versus an appreciative chuckle). In an informal group gathering like a wedding reception, side conversations form, people talk over each other, and there are social rules around eavesdropping that govern when it is appropriate to join a conversation and when to move along. These rich forms of human interaction need more than intelligibility; most crucially they need a sense of distance and directionality. For example, a virtual party is nearly impossible when all voices are merged and rendered with equal volume through a single loudspeaker atop a table. Nor is it easy to replicate the spontaneous pre- and post-meeting discussions that occur as people shuffle in and out of a conference room.

In the remainder of this section, we discuss aspects of audio capture and rendering that better enable a sense of immersion. How is a desired audio field created in an environment? How is audio information extracted from an environment? Perhaps most importantly, exactly what audio field is desirable, especially when coupled with video?

A. Audio Rendering

An audio experience is created, for those with normal hearing, by establishing a sound pressure field in each ear canal. (Low frequencies felt by the whole body are less important for human communication.) This can be achieved, for example, by using a stereo headset, by positioning an array of transducers in a hemisphere around a person's head, or by using one or more transducers positioned some distance from the listener.

The shortcomings of the naive use of a stereo headset are familiar: sounds often seem to originate from inside the skull of the listener, and their apparent position moves inappropriately when the listener's head moves. Yet stereo headphones can do far better. For those listeners familiar only with the limited experiences of today's portable music players, online recordings (see, e.g., [69]) are an effective demonstration of how much more spatial richness is readily achievable. Further, by tailoring the sound to the listener's ear and head shape, and by tracking the position and orientation of the listener's head and re-rendering the sound field appropriately, a stereo headset is capable at least in principle of delivering a fully immersive experience [68].

While possible in principle, real-time rendering of spatially consistent binaural audio for headsets has proved difficult. Sound interacts in important ways with the outer ear, head, torso, and the surrounding room. Every ear is shaped differently, and people shift their head and body to sample the sound field from different poses. The human auditory system is especially sensitive to and in fact relies on these static and dynamic interactions. An aspirational goal in binaural sound reproduction is to compute in real time, with sufficient fidelity to fool a human listener, a sound pressure field in the ear canal that corresponds to a sound source with arbitrary direction and range, ideally without resorting to extensive measurements of listeners, rooms, and transducers. Psychoacoustics plays a key role in current approaches; human insensitivity to certain acoustic details allows many modeling shortcuts to be taken while preserving an illusion of directionality. See [70] for a deeper discussion and a survey of results.

Specialized audio capture solutions can partly sidestep this challenging problem. For example, to immerse a remote participant into a wedding reception, one could use a mobile robot avatar equipped with a synthetic head and stereo microphones. For improved results, the head can be sonically matched to the target listener and can track the listener's head pose in real time [71].

Apart from headphones, sound fields can also be created via one or more stationary loudspeakers. The key challenge remains: what do we emit from the loudspeakers to induce a desired sound field in the ear canals of the listeners? If we constrain the listener's position, apply appropriate sound treatment to the room and furniture, and limit the sound locations we attempt to synthesize, a well-engineered stereo or three-way loudspeaker system can be remarkably effective. This is exactly the situation in many high-end telepresence systems. Relaxing these constraints makes the rendering problem harder, potentially even infeasible.

Given a sufficient number of loudspeakers, coupled with sufficiently accurate models for the room and its occupants, including tracking of listener head positions, one can in principle deliver completely different sounds to every eardrum in the room. In ideal circumstances it suffices to have just one loudspeaker per eardrum. Mathematically, if $H_{ij}(\omega)$ is the transfer function from loudspeaker $i \in 1, \ldots, N$ to eardrum $j \in 1, \ldots, M$, by assembling the transfer functions into an $N \times M$ matrix and computing its inverse, we can calculate the N loudspeaker excitation signals needed to deliver the desired sound signals to the

M target eardrums. Many challenges must be overcome to implement this idea in practice: the transfer functions are difficult to determine with sufficient accuracy, especially in reverberant rooms and at high frequencies; the transfer functions can change quickly due to head movements or because the occupants' movements alter the room acoustics; inversion can be ill-conditioned, which can drive the speakers past linearity; and the signal processing demands a low-latency implementation. The single-listener implementations of [72] and [73] illustrate the progress made over the last decade. The jump from a single- to a many-listener system, particularly in a poorly controlled environment like a typical meeting room, appears at present to be singularly challenging.

B. Audio Capture

The goal of immersive audio capture is the identification and capture of sound field elements that are relevant for remote rendering. One broad category of capture methods isolates clean versions of individual sound sources, such as the individual voices of active talkers, from other interfering sound sources, in particular ambient noise, reverberation, and remote sounds rendered locally by loudspeakers. This can be achieved by mechanical and algorithmic methods, including close talking microphones, microphone arrays, and blind source separation techniques. In many cases, local reverberations are undesirable at the remote site; removing reverberation algorithmically remains an active area of research. Remote sounds rendered by local loudspeakers lead to the well-known and hard problem of *echo cancellation*, which is endemic to two-way audio communication. Immersive rendering systems that employ loudspeakers must thus solve two coupled difficult problems: how to induce desired sounds in the ears of the listeners, and how to remove the resulting contamination from the microphone signals.

A second broad category of capture methods does not attempt to isolate individual sounds, but rather captures a sound field at a location or over an area, for example, at the ears of a virtual listener, around the head of a virtual listener, or across a virtual window. The sound field is then conveyed in its entirety to a remote location. An advantage of this approach is its ability to provide all the ambient sounds in an environment, not just human speech. It also sidesteps many of the difficult algorithmic issues described above, though not echo cancellation. Disadvantages include the need for a potentially high network bandwidth, the difficulty in synthesizing a virtual geometry different from that of the capture location, and the difficulty in knitting together captured sound fields from multiple remote locations simultaneously. Certain high-end telepresence rooms today adopt a highly simplified but effective version of this capture method: three microphones in a local room separately drive three correspondingly positioned speakers in a remote location. This approach is effective at providing a natural form of spatial audio for a collaborative meeting between two rooms with several people in each room. Specifically, this helps provide the necessary audio/video correspondence, described next.

C. Audio/Video Correspondence

Video is powerful cue for audio, particularly when the video is perceived as "relevant" to the audio, such as a talking head [74]. When confronted with a conflict between visual position and audio position, even a trained listener will fuse the audio to the video position for a separation up to 10°. An untrained listener will tolerate up to 20°. Thus it would seem that one could relax some constraints on audio rendering when video is present, an effect that is relied upon in movies. Indeed, in a communication setting, if the remote side of a collaboration session is portrayed on a display that subtends less than 40° of view (which is about as close as most people get to a screen today), for untrained listeners it may be acceptable to place all spoken audio in the middle of the display.

The simplification above likely breaks down when there are multiple remote parties talking, or when the ambient sounds are an important part of the communication. Consider the case of a family dinner in which the two halves of the dinner table are separated by a thousand miles with a display in between. There will be multiple simultaneous talkers—at least in many families—and the ambient sounds are an important part of the social impact of the event. Audio should be good enough in this situation to allow the "cocktail party effect" to take place [75], [76].

Audio and video interact strongly with respect to quality; it is well known that good audio makes people feel that the video is better. Also, nonsynchronized audio and video, resulting in lack of lip-sync, may cause a reduced overall experience.

The requirement for immersive audio rendering becomes more acute when the listener is surrounded on several sides by collaboration screens, or when the video portrays a virtual geometry with a more complex structure. For example, in Second Life [77], there is the sense that the world continues left, right, and behind what can be seen on the display, hence there is a need to render sounds from positions that are not currently visible.

If we have a video rendering system that can place a virtual Yoda [78] on a physical chair, then we want his voice to follow him. But in the absence of an augmented reality representation of a person, do we want audio to arrive as a disembodied voice floating over an otherwise empty chair? Maybe so, maybe not—what seems socially awkward today could be viewed as entirely natural in the future.

What happens when we combine two wildly disparate acoustic environments, like a living room and a train station? Do we want the noise and echoing openness of the train station to flood in? The answer may depend on the

context and purpose of the communication; a business meeting will likely have different requirements than the case of a family keeping an at-home relative immersed in an ongoing vacation.

VI. COMMUNICATION ASPECTS

Multimedia communication has been made possible today by massive improvements in our ability to compress multimedia signals coupled with ever increasing network capacity. Better compression has arisen from a better understanding of compression principles as well as from enormous increases in available computation and memory. The striking improvements in computation in recent decades help suggest the scale of potential future improvements. When HDTV was being developed in 1991 it took 24 h to compress 1 s of 720P HDTV video on a high-end workstation; in 2011, this can be done in real time on a smartphone you hold in your hand. The improvements arose from faster processors and greater memory bandwidth, as well as instruction-level parallelism, multicore processors, and graphics processing units which can contain hundreds of cores. Similarly, the advent of high bandwidth wired and wireless packet networks, with their flexibility for dynamic resource allocation, has been crucial for transporting these multimedia signals. The future promises additional improvements in computation, storage, and networking, and these are necessary to enable the immersive communication envisioned in this paper. In addition, the future will see more processing being performed in the cloud. This is especially true for mobile devices, which may have limited compute capability or battery life. For example, many applications such as visual search or 3-D scene reconstruction can be partitioned such that select processing is performed on the device (e.g., feature extraction) and the remaining processing in the cloud (e.g., compute-intensive search across large databases).

Video and audio coding has been a research focus in the academic and industry communities for many years and this has resulted in a variety of international standards by the ISO/IEC, ITU, etc., that have enabled interoperable multimedia communication [37]. The increasing desire for interoperability across different end-devices and systems is likely to continue the pressure to develop and deploy standardized solutions for compression, transport, and interaction.

Four likely trends in the future are: 1) increased scale of the raw size for the captured and rendered signals; 2) increased scale for the range of bit rates and diversity of endpoints which need to be supported; 3) new sensors which capture different attributes; and 4) new associated representations, compression algorithms, and transport techniques. These four trends are briefly discussed next.

The future will likely see large immersive displays, as described in Section IV-A, with 2-D displays going from their current several megapixels per frame, to tens or hundreds of megapixels per frame (and possibly gigapixels per frame in special cases). Light field displays will require several orders of magnitude larger number of light rays. This scaling of visual information is illustrated in Table 1.

These dramatic increases in the size of the raw visual or auditory information will require new compression techniques to squeeze them within the available network bandwidth and storage capacity in the future. One important characteristic which makes this feasible is that as the sampling rate increases, so does the redundancy between samples, and therefore the compressed bit rate increases much slower than the sampling rate. For example, when increasing the number of light rays by 10× for every spatial location, the redundancy across the light rays results in the compressed data rate increasing by much less than a factor of 10. This is one of the assumptions made in Table 1. Furthermore, a second important characteristic is that the perceptually relevant information also does not scale linearly with sampling rate. High-quality multimedia compression relies on reducing the redundancy and the perceptually irrelevant information within the media signals [54]. The research community has many years of understanding of the redundant and perceptually irrelevant information in 2-D video and audio—the near future will see an increased focus on understanding these issues for immersive visual, auditory, and haptic data. This new knowledge will most likely direct the creation of new multimedia representations and compression algorithms.

The future will likely see a wide diversity of communication endpoints, ranging from custom-designed immersive studios, to systems that can be rolled into an existing room or environment, to desktop and mobile devices.

Table 1 Example of the Large Increase in Raw and Compressed Data Rates for Going From a Conventional Digital TV Video (Row 1) to a Large Display Providing an Immersive Experience (Row 2), and to Future 3-D Light Field Displays (Rows 3 and 4). For Simplicity, 60 Frames/s Is Assumed for all Applications, and 1000 × 1000 Light Rays (Horizontally and Vertically) for Each (x, y) Coordinate

Application	Mpixels/f	Light rays/pixel	M light rays/s	Bit rate
DTV	1	1	60	5 Mb/s
Large DTV	40	1	2,400	200 Mb/s
3D	1	10^6	60,000,000	~ Gb/s
Large 3D	40	10^6	2,400,000,000	Many Gb/s

These endpoints will be able to support, and require, vastly different versions of the multimedia content. For example, an immersive studio may need 100× or 1000× the number of pixels of a desktop or mobile system. Similarly, there will probably be wide diversity in the bandwidth available to different endpoints based on the network they are connected to, and the time-varying demands placed on the network by other applications. These heterogeneous and time-varying requirements provide additional challenges for future systems, which may need to simultaneously communicate in a multiparty scenario with, for example, an immersive studio, a desktop system, and a mobile device.

Today, video and audio coding directly compresses the captured visual or auditory waveforms from 2-D imaging sensors or discrete microphones. This has been a practical, robust, and successful approach for current applications. The future promises a number of significant extensions. Three-dimensional depth cameras capture both the red–green–blue (RGB) colors and depth for every pixel. Plenoptic cameras capture multiple light rays coming from different directions. Holographic capture devices capture the amplitude and phase of the optical wavefront [38]. These, and other, new sensors will provide a more detailed understanding of the visual and auditory scene. To exploit this requires the development of new compression and transport techniques to both efficiently deliver the content and leverage the new understanding to provide new capabilities to modify or enhance the scene.

The future is likely to see the successful development of novel multimedia representations and associated compression algorithms. Video compression for stereo and multiview video coding has recently been standardized [37], and ongoing efforts are incorporating depth maps in the compression. Distributed coding principles have also been applied for both independent coding of the video from multiple cameras with joint decoding, and for creating low-complexity encoders [40]. In the future, the visual information may be expressed as a collection of 2-D layers or 3-D objects, lighting information, and a scene description of how to compose them to render the scene [37], [42]. Similarly, the auditory information may be expressed as a collection of audio sources (e.g., individual people) and environmental effects such as reverberation, and a scene description of how to compose them to render the auditory scene [43]. Object-based video and audio coding was developed within the MPEG-4 standard in the late 1990s, however it was not successful because of the difficulty of decomposing a scene into objects, and the high computational requirements. For example, segmenting video into meaningful objects is a very challenging problem. Fortunately, the advent of 3-D-depth cameras provides a major step forward. Similarly, algorithms for determining the number of speakers in a room and separating their voices are improving. Since immersive communication involves both video and audio, multimodal processing can help, where, for example, face detection and tracking is applied to help estimate the number of speakers and their bearing relative to a microphone array, and thereby the visual information can guide the audio processing. These object-based systems would analyze the captured signals, decompose them into meaningful objects, and appropriately compress and deliver each of them. The separate coding and transport of the different objects greatly facilitates object-based processing, such as the addition or removal of objects, or the placement of objects within a virtual visual or auditory environment. Another trend is toward model-based image/video synthesis, where textures such as hair or grass are created which are conceptually faithful, though not pixel-wise accurate. Future collaborative meetings may also have some participants photo-realistically rendered, and others expressed as avatars because of limited bandwidth, lack of available camera, or privacy preferences. The compression technique selected for a specific immersive communication session would depend on the available bandwidth, computational capability, available sensors, and application-level constraints such as privacy concerns.

Immersion in a remote environment requires the ability to physically interact with objects in that environment. This relies on the haptics modality, comprising tactile sense (touch, pressure, temperature, pain) and kinaesthetics (perception of muscle movements and joint positions). To provide timely haptic feedback requires a control loop with roughly 1-ms latency. This corresponds to a high packet transmission rate of nominally 1000 packets/s. To overcome this problem, recent work leveraged perceptual models to account for the human haptic sensitivity. For example, based on Weber's law of the just noticeable difference (JND), the JND is linearly related to the intensity of the stimulus. Changes less than the JND would not be perceived and hence do not need to be transmitted. This leads to the notion of perceptual dead-bands within which changes to the haptic signal would in principle be imperceptible to the user [44]. Accounting for haptic signal perceptibility, coupled with conventional compression techniques, can lead to significant reductions in the transmitted packet rate and the total bit rate transmitted [45]. Haptics, and the compression and low-delay two-way transmission of haptics information, is an emerging area that promises to dramatically increase the realism and range of uses of immersive communication systems.

VII. QUALITY OF EXPERIENCE AND HUMAN PERCEPTION

Ultimately, the success or failure of any system for immersive communication lies in the quality of the human experience that it provides, not in the technology that it uses.

As mentioned in the Introduction, immersive communication is about exchanging natural social signals between remote people and/or experiencing remote locations in ways that suspend one's disbelief in being there. Fundamentally, the quality of such an immersive experience

Fig. 5. ***HP Halo (top) and Cisco Telepresence (bottom).***

relates to human perception and mental state. For immersive communication, while it may be sufficient to reproduce the sensory field at a distant location, it may not be necessary, or even feasible, to do so. More important to achieving a high QoE is inducing in each participant an intended illusion or mental state.

Hence, it may come as no surprise that the latest generation of immersive communication products, which set a new standard for QoE, was conceived by veteran storytellers in Hollywood. Moviemaker DreamWorks Animation SKG, in partnership with technologist Hewlett-Packard, developed the HP Halo Video Collaboration Studio, shown in Fig. 5 bringing a moviemaker's perspective to video conferencing. (Various companies, including Cisco, Tandberg, Teliris, and Polycom, have since offered similar systems.) Unlike their predecessor video conferencing systems, these high-end telepresence systems are carefully designed to induce participants to suspend their disbelief in being with each other in the same room. In fact these high-end telepresence systems preserve many of the important immersive cues listed in Table 2, auditory and visual, beginning with peripheral awareness and consistency. Peripheral awareness is the awareness of everything that is going on in one's immediate vicinity, providing a sense of transparency, of being there. Halo achieves that, in part, by a large field of view. Consistency between local and remote sites is important for maintaining the illusion of a single room. So all elements of room design, such as tables, chairs, colors, wall coverings, acoustics, and lighting, match across all locations by design. In addition, a number of measures were taken to maintain consistent 3-D cues. For example:

- remote people appear life size and occupy the same field-of-view as if they were present;
- multichannel audio is spatially consistent with the video;
- video streams from remote sites are placed in an order that is consistent with a virtual meeting layout, improving eye contact and maintaining consistent gaze direction.

In addition:

- low latency and full-duplex echo-controlled audio allow natural conversation;
- people's facial expressions are accurately conveyed by a combination of sufficient video resolution and constrained seating that keeps participants in front of the cameras.

Despite the attention paid to these details, other conflicting cues can serve to break the illusion. One example is eye contact. In a real physical space, each person is able to make eye contact with at most one other person. But in today's telepresence systems, eye contact is compromised because all local participants see the same view of each remote participant. Future multiview 3-D displays promise to improve on this aspect.

Evaluating the QoE of an immersive communication system can be done on several levels, which one could call

Table 2 Immersive Cues

	Auditory	Visual
Peripheral Awareness	Peripheral audio	Peripheral video
Consistency	E.g., reverberation	E.g., lighting
Spatial cues	Relative loudness	Relative size
	Direction/HRTF	Perspective
	Direct/Reflected	Occlusion, shadow
	Direct/Reverberant	Shading, color
	Reverberation	Haze, contrast
	Occlusion	Stereo parallax
	Absorption	Motion parallax
	Reflection	Focus

the performance, psychometric, and psychophysical levels. At the performance level, QoE is evaluated in terms of the overall effectiveness of the experience with respect to performing some task. For example, Bailenson *et al.* [46] studied whether gaze awareness could reduce the number of questions, time per question, or completion time in a "game of 20 questions" with multiple players. Such evaluations are often performed as user studies in a lab, since well-defined, measurable tasks must be carried out. Clearly, measures of performance may be completely different for one task (e.g., playing poker) compared to another (e.g., negotiating a business deal). Hence, it is possible for different immersive communication systems to excel in different settings.

At the psychometric level, QoE is evaluated in terms of how the participants feel about the experience. Such evaluations can be carried out either as a study in a lab (often in the context of performing a task) or as a field study (in the context of a regular workload). As an example of a field study, Venolia *et al.* [47] studied the deployment of embodied social proxies, or telepresence robots, in four real-world software development teams in which one member of each team was remote. Within each team, an embodied social proxy for the remote team member was deployed for six weeks. A baseline survey was completed before the deployment, and the same survey was completed after the deployment, by all team members. The survey listed a set of assertions to be scored on an eight-point range from 0 (low agreement) to 7 (high), known as a Likert scale in the psychometric literature [48]: one set of assertions related to meeting effectiveness (e.g., "I think X has a good sense of my reactions"); another set of assertions related to awareness (e.g., "I am aware of what X is currently working on and what is important to X"); and a final set of assertions related to social aspects (e.g., "I have a sense of closeness to X"). Comparing the surveys before and after the deployment yielded quantitative psychometric results on the overall experience with respect to meeting effectiveness, awareness, and social aspects. Qualitative results were also obtained through freeform comments on the surveys as well as direct observations of the teams in action by ethnographers.

Other examples of psychometric evaluations include those of [49]–[51], which showed the importance of awareness of remote participants' object of attention on collaborative tasks, and [52], which showed that trust can be improved using a multiview display to support proper eye gaze. These were assessed in lab studies using surveys.

The third level on which to evaluate QoE is the psychophysical level. Psychophysics is the study of the relation between stimuli and sensation. Weber's law, that the threshold of perception of change in a stimulus is proportional to the amplitude of the stimulus (resulting in a logarithmic relationship between a stimulus' amplitude and its perception), is a well-known psychophysical result. Psychophysical evaluations of the QoE of an immersive communication system are based on subjective perception of a set of stimuli provided by the system. Psychophysical studies of audio and visual quality have been well studied through mean opinion scores (MOS) [53], JNDs due to quantization noise [54], frequency sensitivity, and so forth. Rules of thumb for high-frequency sensitivity are that human hearing falls off beyond 15–20 kHz [55] and visual acuity in the fovea falls off beyond 60 cycles per degree [56]. There are also well-known results on delay, such as interactive voice conversations can tolerate up to 150 ms of delay [57], and that for lip sync, audio can lag video by up to 100–125 ms, but can lead video by up to only 25–45 ms [58]. Interestingly, recent evidence suggests that video tends to mask audio delay, and that interactive *audiovisual* conversations can tolerate delays significantly larger than 150 ms [59], the hypothesis being that intent to speak can be signaled through video instead of audio. Eye gaze has also been studied: eye contact can be perceived (at 90th percentile) within an error of 7° downward, but only 1° left, right, or upward [60].

However, many other psychophysical characterizations of QoE for immersive communication are incipient. One area that is important to immersive communication is quality of 3-D modeling [61] and stereoscopy [62], which are largely still art forms. We are only beginning to understand how humans perceive mismatches between the various cues in Table 2, such as any mismatch between the spatially perceived source of audio and its visual counterpart [63]; between visual stereo parallax, motion parallax, and focus [64]; between the lighting, color, or acoustics of local and remote sites; etc. And as of yet we have no studies of the tolerable delay between head motion and the change of visual field for motion parallax. We have little understanding of the fidelity of geometric space, for example, how much it can be warped before one perceives changes in gesture. One thing is clear: people can watch video on flat screens of any size at a wide range of orientations, and perceive the scene from the camera's point of view, easily accommodating changes in the camera's position and focal length. There is an analogous robustness to color balance. These factors point to remarkable "constancy" properties of human perception [65] whose role in immersive communication has yet to be understood. Overall, some elements of an immersive experience may be quite flexible, while others must match physical reality, or else risk user fatigue that threatens long-term use.

Finally, it must be noted that human psychology can play an unexpectedly critical role in determining QoE. This is evident, for example, in our ability to detect when the representation of a human has entered an "uncanny valley" [66] and in our tendency to imbue devices with anthropomorphic properties [66], [67].

VIII. CONCLUDING REMARKS

After many decades of steady improvements, underlying technical trends (e.g., in computation, bandwidth, and

resolution) are rapidly accelerating progress towards new and more immersive communication systems. Indeed, we are experiencing a blossoming variety of immersive communication systems beyond standard audio and video, as outlined in Section III.

Moreover, as technology pushes (i.e., supplies), society pulls (i.e., demands). The societal need for more effective ways for remote people to communicate with each other, as if they were face to face, has never been greater.

Numerous societal conditions are aligning to create this need: the need to reduce environmental impact, the need to increase economic efficiency through reduced travel costs and fatigue (both in terms of fees and wasted time), the need for richer collaboration in an ever more complex business environment, and the difficulty or undesirability of travel during natural disasters such as volcanic eruptions and influenza and influenza outbreaks [79], [80]. Some of the foremost in importance relate to climate, energy, and the environment. Many of us travel internationally several times per year to attend conferences and other events. Each time we do so, we consume several barrels of oil and release on the order of a ton of CO_2 into the atmosphere, per person, contributing to global warming at an unprecedented scale.

The economy and productivity are not far behind in importance. Whether an economy is booming or busting, it is becoming more and more critical to connect people with jobs and businesses with workers, wherever they may be around the world. The possibilities for immersive communication to address this need are ripe. Studies have shown that information workers around the world spend over 60% of their time in some form of communication with other people. Furthermore, over 60% of information workers report that they could fulfill their job duties from a remote location if they were permitted to do so [79]. Moreover, the time saved commuting would be substantial (in the United States, 45-min average daily commute: almost 10% of overall productivity). These are gains we cannot afford to ignore.

In a sign of the times, physical security has also risen to the top of needs addressable by more immersive communication. When Iceland's Eyjafjallajökull volcano erupted in April 2010, clouds of ash ended up canceling over 100 000 flights, stranding over eight million passengers in Europe, and costing the airline industry over 2.5 billion euros. Similarly costly disruptions accompanied hurricanes in the United States, earthquakes and tsunamis in Japan, and bird flu in Asia, not to mention acts of terrorism. The economic impact of such threats could be reduced by improved forms of communication, which could allow most business to be carried on as usual.

Then, of course, there are the ever-present needs to bring families closer together and to improve the availability of quality of education and healthcare around the world.

In short, the technology is advancing and the social need is there. In this paper, we have tried to put technology, applications, and scenarios in a historical perspective with an eye on promising future technology directions. Furthermore, we have tried to emphasize the importance of evaluating the experience from the human user's perspective. Some of the most important future advances will come from ingenious combinations of core components (e.g., light field displays and 3-D depth sensors) in unanticipated ways. We are confident that we will see substantial progress over the next decade on the road to immersive communication. ■

REFERENCES

[1] T. Friedman, *The World Is Flat: A Brief History of the Twenty-First Century*, 3rd ed. New York: Farrar, Straus and Giroux, 2007.

[2] T. Clark, *Faraway, So Close: Doing Global iPhone Dev from New Zealand*, Jan. 2011. [Online]. Available: http://arstechnica.com/gaming/2011/01/sell-in-america-live-in-new-zealand.ars

[3] J. N. Bailenson, A. C. Beall, and J. Blascovich, "Mutual gaze and task performance in shared virtual environments," *Personality Social Psychol. Bull.*, pp. 1–15, 2003.

[4] A. Pentland, *Honest Signals: How They Shape Our World*. Cambridge, MA: MIT Press, 2008.

[5] A. Vinciarelli, M. Pantic, and H. Bourland, "Social signal processing: Survey of an emerging domain," *J. Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[6] M. Zimmerman, "The nervous system in the context of information theory," in *Human Physiology*, R. F. Schmidt and G. Thews, Eds. New York: Springer-Verlag, 1989, pp. 166–173.

[7] B. Hoang and O. Perez, *Terabit Networks*. [Online]. Available: http://www.ieee.org/portal/site/emergingtech/techindex.jsp?techId=1140

[8] M. Minsky, "Telepresence," *OMNI Mag.*, pp. 44–52, Jun. 1980.

[9] G. du Maurier, "Edison's telephonoscope (transmits light as well as sound)," *Punch's Almanack*, Dec. 9, 1878.

[10] A. G. Bell, "On the possibility of seeing by electricity," in *Alexander Graham Bell Family Papers*. Washington, DC: Library of Congress, Apr. 10, 1891.

[11] I. E. Sutherland, "The ultimate display," in *Proc. IFIP Congr.*, 1965, pp. 506–508.

[12] "S.F. man's invention to revolutionize television," *San Francisco Chronicle*, vol. CXXXIII, no. 50, Sep. 3, 1928.

[13] C. Cruz-Niera, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, "The CAVE: Audio visual experience automatic virtual environment," *Commun. ACM*, vol. 35, no. 6, Jun. 1992, DOI: 10.1145/129888.129892.

[14] J. Fihlo, K. M. Inkpen, and M. Czerwinski, "Image, appearance and vanity in the use of media spaces and videoconference systems," *Proc. ACM Int. Conf. Supporting Group Work*, May 2009, DOI: 10.1145/1531674.1531712.

[15] D. Lindbergh, "The past, present, and future of video telephony: A systems perspective," Keynote, Packet Video Workshop, Dec. 2004.

[16] N. Jouppi, S. Iyer, S. Thomas, and A. Slayden, "BiReality: Mutually-immersive telepresence," *Proc. ACM Int. Conf. Multimedia*, 2004, DOI: 10.1145/1027527.1027725.

[17] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computation Models of Visual Processing*, M. Landy and J. A. Movshon, Eds. Cambridge, MA: MIT Press, 1991, pp. 3–20.

[18] A. Jeans, M. Almanza-Workman, R. Cobene, R. Elder, R. Garcia, R. F. Gomez-Pancorbo, W. Jackson, M. Jam, H.-J. Him, O. Kwon, H. Luo, J. Maltabes, P. Mei, C. Perlov, M. Smith, and C. Taussig, "Advances in roll-to-roll imprint lithography for display applications," in *Proc. SPIE*, Feb. 2010, vol. 7637, p. 763719.

[19] H. J. Kim, M. Almanza-Workman, B. Garcia, O. Kwon, F. Jeffrey, S. Braymen, J. Hauschildt, K. Junge, D. Larson, D. Stieler, A. Chaiken, B. Cobene, R. Elder, W. Jackson, M. Jam, A. Jeans, H. Luo, P. Mei, C. Perlov, and C. Taussig, "Roll-to-roll manufacturing of electronics on flexible substrates using self-aligned imprint lithography (SAIL)," *J. Soc. Inf. Display*, vol. 17, p. 963, 2009, DOI: 10.1889/JSID17.11.963.

[20] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, and P. Debevec, "Achieving eye contact in a one-to-many 3D video teleconferencing system," *ACM Trans. Graphics*, vol. 28, no. 3, Jul. 2009, DOI: 10.1145/1531326.1531370.

[21] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.

[22] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Stanford Univ. Comput. Sci., Stanford, CA, Tech Rep. CSTR 2005-02, Apr. 2005.

[23] M. Levoy and P. Hanrahan, "Light field rendering," *Proc. ACM SIGGRAPH*, 1996, pp. 31–42.

[24] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen, "The lumigraph," *Proc. ACM SIGGRAPH*, 1996, pp. 43–54.

[25] C. Zhang and T. Chen, "A survey on image-based rendering-representation, sampling and compression," in *EURASIP Signal Process., Image Commun.*, 2004, vol. 19, pp. 1–28.

[26] J. X. Chai, S. C. Chan, H. Y. Shum, and X. Tong, "Plenoptic sampling," in *Proc. SIGGRAPH*, 2000, pp. 307–318.

[27] C. Zhang and T. Chen, "Spectral analysis for sampling image-based rendering data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 11, pp. 1038–1050, Nov. 2003.

[28] M. N. Do, D. Marchand-Maillet, and M. Vetterli, "On the bandwidth of the plenoptic function," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 708–717, Feb. 2012.

[29] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, "A state of the art in structured light patterns for surface profileometry," *Pattern Recognit.*, vol. 43, no. 8, pp. 2666–2680, Aug. 2010.

[30] Q. Yang, K. H. Tan, B. Culbertson, and J. Apostolopoulos, "Fusion of active and passive sensors for fast 3D capture," *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2010, pp. 69–74.

[31] A. Kolb, E. Barth, and R. Koch, "ToF-sensors: New dimensions for realism and interactivity," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, DOI: 10.1109/CVPRW.2008.4563159.

[32] H. Urey, K. V. Chellappan, E. Erden, and P. Surman, "State of the art in stereoscopic and autostereeoscopic displays," *Proc. IEEE*, vol. 99, no. 4, pp. 540–555, Apr. 2011.

[33] K.-H. Tan, I. Robinson, R. Samadani, B. Lee, D. Gelb, A. Vorbau, B. Culbertson, and J. Apostolopoulos, "Connectboard: A remote collaboration system that supports gaze-aware interaction and sharing," *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2009, DOI: 10.1109/MMSP.2009.5293268.

[34] S. C. Chan, H. Y. Shum, and K. T. Ng, "Image-based rendering and synthesis," *IEEE Signal Process. Mag.*, vol. 24, no. 7, pp. 22–33, Nov. 2007.

[35] R. Szeliski, *Computer Vision: Algorithms and Applications*. New York: Springer-Verlag, 2011.

[36] M. Hirsch, D. Lanman, H. Holtzman, and R. Raskar, "BiDi screen: A thin, depth-sensing LCD for 3D interaction using light fields," *ACM Trans. Graphics*, vol. 28, no. 5, 2009.

[37] L. Chiariglione, "Multimedia standards: Interfaces to innovation," *Proc. IEEE*, to be published.

[38] V. M. Bove, "Display holography's digital second act," *Proc. IEEE*, vol. 100, no. 4, Apr. 2012, DOI: 10.1109/JPROC.2011.2182071.

[39] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," *Proc. IEEE*, vol. 99, no. 4, pp. 626–642, Apr. 2011.

[40] B. Girod, A. M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[41] J. Y. A. Wang and E. Adelson, "Representing moving images with layers," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 625–638, Sep. 1994.

[42] A. Smolic and P. Kauff, "Interactive 3D video representations and coding technologies," *Proc. IEEE*, vol. 93, no. 1, pp. 98–110, Jan. 2005.

[43] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured audio: Creation, transmission, and rendering of parametric sound representations," *Proc. IEEE*, vol. 86, no. 5, pp. 922–940, May 1998.

[44] P. Hinterseer, S. Hirche, S. Chaudhuri, E. Steinbach, and M. Buss, "Perception-based data reduction and transmission of haptic data in telepresence and teleaction systems," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 588–597, Feb. 2008.

[45] E. Steinbach, S. Hirche, J. Kammerl, I. Vittorias, and R. Chaudhari, "Haptic data compression and communication," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 87–96, Jan. 2011.

[46] J. N. Bailenson, A. C. Beall, and J. Blascovich, "Gaze and task performance in shared virtual environments," *J. Vis. Comput. Animat.*, vol. 13, no. 5, pp. 313–320, 2002.

[47] G. Venolia, J. Tang, R. Cervantes, S. Bly, G. Robertson, B. Lee, K. Inkpen, and S. Drucker, "Embodied social proxy: Connecting hub-and-satellite teams," *Proc. Comput. Supported Cooperative Work*, 2010.

[48] P. E. Spector, *Summated Rating Scale Construction: An Introduction*. Thousand Oaks, CA: Sage Publications, 1992.

[49] J. Tang and S. Minneman, "VideoWhiteboard: Video shadows to support remote collaboration," *Proc. SIGHI Conf. Human Factors Comput. Syst.*, 1991, pp. 315–322, DOI: 10.1145/108844.108932.

[50] H. Ishii and M. Kobayashi, "ClearBoard: A seamless medium for shared drawing and conversation with eye contact," *Proc. SIGHI Conf. Human Factors Comput. Syst.*, 1992, DOI: 10.1145/142750.142977.

[51] A. Tang, M. Pahud, K. Inkpen, H. Benko, J. C. Tang, and B. Buxton, "Three's company: Understanding communication channels in a three-way distributed collaboration," *Proc. Comput. Supported Cooperative Work*, 2010, pp. 271–280.

[52] D. T. Nguyen and J. Canny, "Multiview: Improving trust in group video conferencing through spatial faithfulness," *Proc. SIGHI Conf. Human Factors Comput. Syst.*, 2007, DOI: 10.1145/1240624.1240846.

[53] *Methods for Objective and Subjective Assessment of Quality*, ITU-T Recommendation P.800, 1996.

[54] N. S. Jayant, J. D. Johnston, and R. J. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422, Oct. 1993.

[55] B. C. J. Moore, *Psychology of Hearing*, 5th ed. New York: Academic, 2003.

[56] M. H. Pirenne, "Visual acuity," in *The Eye*, vol. 2, H. Davson, Ed. New York: Academic, 1962.

[57] *One-Way Transmission Time*, ITU-T Recommendation G.114, 2003.

[58] *Relative Timing of Sound and Vision for Broadcasting*, ITU-R BT.1359-1, 1998.

[59] E. Geelhoed, A. Parker, D. J. Williams, and M. Groen, "Effects of latency on telepresence," HP Labs, Tech. Rep. HPL-2009-120, 2009.

[60] M. Chen, "Leveraging the asymmetric sensitivity of eye contact for videoconferencing," *Proc. SIGHI Conf. Human Factors Comput. Syst.*, 2002, DOI: 10.1145/503376.503386.

[61] I. Cheng, R. Shen, X.-D. Yang, and P. Boulanger, "Perceptual analysis of level-of-detail: The JND approach," *Proc. Int. Symp. Multimedia*, 2006, pp. 533–540.

[62] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP J. Image Video Process.*, vol. 2008, 2008, DOI:10.1155/2008/659024.

[63] S. Boyne, N. Pavlovic, R. Kilgore, and M. H. Chignell, "Auditory and visual facilitation: Cross-modal fusion of information in multi-modal displays," in *Meeting Proc. - Visualisation Common Operational Picture.*, Neuilly-sur-Seine, France, pp. 19-1–19-4, 2005, Paper 19, RTO-MP-IST-043. [Online]. Available: http://www.rto.nato.int/abstracts.asp

[64] M. Lambooij, W. IJsselsteijn, and I. Heynderickx, "Stereoscopic displays and visual comfort: A review," in *SPIE Newsroom*, Apr. 2, 2007, DOI: 10.1117/2.1200703.0648.

[65] V. Walsh and J. Kulikowski, Eds., *Perceptual Constancy: Why Things Look as They Do*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[66] J. Blascovich and J. Bailenson, *Infinite Reality: Avatars, Eternal Life, New Worlds, and the Dawn of the Virtual Revolution*. New York: Harper-Collins, 2011.

[67] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Stanford, CA: CSLI Publications, 1996.

[68] QSound Labs, *Virtual Barber Shop*, May 5, 2011. [Online]. Available: http://www.qsound.com/demos/binaural-audio.htm

[69] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 33–42, Jan. 2011.

[70] B. Kapralos, M. R. Jenkin, and E. Milios, "Virtual audio systems," *Presence*, vol. 17, no. 6, pp. 527–549, 2008.

[71] I. Toshima, S. Aoki, and T. Hirahara, "An acoustical tele-presence robot: TeleHead II," in *Proc. IEEE/RSJ Int. Conf. Intel. Robots Syst.*, Sendai, Japan, Sep. 2004, pp. 2105–2110.

[72] W. Gardner, "3-D audio using loudspeakers," Ph.D. dissertation, Progr. Media Arts Sci., Schl. Archit. Planning, Massachusetts Inst. Technol., Cambridge, MA, 1997.

[73] T. Grämer, "Efficient modeling of head movements and dynamic scenes in virtual

acoustics," M.S. thesis, Univ. Zurich/ETH, Faculty Med., 2010.

[74] D. R. Begault, "Auditory and non-auditory factors that potentially influence virtual acoustic imagery," in *Proc. 16th Audio Eng. Soc. Int. Conf. Spatial Sound Reproduction*, Mar. 1999.

[75] A. R. A. Conway, N. Cowan, and M. F. Bunting, "The cocktail party phenomenon revisited: The importance of working memory capacity," *Psychonomic Bull. Rev.*, vol. 8, no. 2, pp. 331–335, 2001.

[76] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005.

[77] Second Life. [Online]. Available: http://www.secondlife.com

[78] Yoda. [Online]. Available: http://en.wikipedia.org/wiki/Yoda

[79] U.S. National Remote Working Survey, 2010. [Online]. Available: http://www.microsoft.com/presspass/download/features/2010/NationalRemoteWorkingSummary.pdf

[80] "Iceland volcano ash hits European flights again," *Agence France-Presse*, May 8, 2010. [Online]. Available: http://news.ph.msn.com/top-stories/article.aspx?cp-documentid=4076015; http://www.google.com/hosted-news/afp/article/ALeqM5gBr_mBjSK4JC-W3iODp3Hhw5S5SiQ

ABOUT THE AUTHORS

John G. Apostolopoulos (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Massachusetts Institute of Technology (MIT), Cambridge.

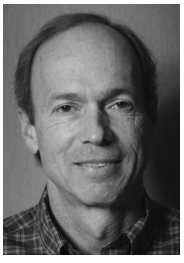
He is Director of the Mobile & Immersive Experience Lab, within Hewlett-Packard Laboratories, Palo Alto, CA, directing about 75 researchers in the United States, India, and the United Kingdom. His research interests include mobile and immersive multimedia communications and improving their naturalness, reliability, fidelity, scalability, and security. In graduate school, he worked on the U.S. Digital TV standard and received an Emmy Award Certificate for his contributions. He enjoys teaching and conducts joint research at Stanford University, Stanford, CA, where he was a Consulting Associate Professor of Electrical Engineering (2000–2009), and is a frequent visiting lecturer at MIT's Department of Electrical Engineering and Computer Science.

Dr. Apostolopoulos received a best student paper award for part of his Ph.D. dissertation, the Young Investigator Award (best paper award) at the 2001 Visual Communication and Image Processing (VCIP 2001) Conference for his work on multiple description video coding and path diversity, was named "one of the world's top 100 young (under 35) innovators in science and technology" (TR100) by *MIT Technology Review*, and was coauthor for the best paper award at the 2006 International Conference on Multimedia and Expo (ICME 2006) on authentication for streaming media. His work on secure transcoding was adopted by the JPSEC standard. He served as chair of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee, and member of Multimedia Signal Processing Technical Committee, technical co-chair for the 2007 IEEE International Conference on Image Processing (ICIP'07), and recently was co-guest editor of the special issue of the IEEE SIGNAL PROCESSING MAGAZINE on "Immersive Communication." He serves as technical co-chair of the 2011 IEEE International Conference on Multimedia Signal Processing (MMSP'11) and the IEEE 2012 International Conference on Emerging Signal Processing Applications (ESPA'12) (and co-founder).

Philip A. Chou (Fellow, IEEE) received the B.S.E. degree from Princeton University, Princeton, NJ, in 1980 and the M.S. degree from the University of California Berkeley, Berkeley, in 1983, both in electrical engineering and computer science, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1988.

From 1988 to 1990, he was a member of Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ. From 1990 to 1996, he was a Member of Research Staff at the Xerox Palo Alto Research Center, Palo Alto, CA. In 1997, he was manager of the compression group at VXtreme, an Internet video startup in Mountain View, CA, before it was acquired by Microsoft in 1997. From 1998 to the present, he has been a Principal Researcher with Microsoft Research in Redmond, WA, where he currently manages the Communication and Collaboration Systems research group. He has served as Consulting Associate Professor at Stanford University (1994–1995), Affiliate Associate Professor at the University of Washington (1998–2009), and Adjunct Professor at the Chinese University of Hong Kong (since 2006). He has longstanding research interests in data compression, signal processing, information theory, communications, and pattern recognition, with applications to video, images, audio, speech, and documents. He is co-editor, with M. van der Schaar, of the book *Multimedia over IP and Wireless Networks* (Amsterdam, The Netherlands, Elsevier, 2007).

Dr. Chou served as an Associate Editor in source coding for the IEEE TRANSACTIONS ON INFORMATION THEORY from 1998 to 2001, as a Guest Editor for special issues in the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), and the IEEE SIGNAL PROCESSING MAGAZINE in 1996, 2004, and 2011, respectively. He was a member of the IEEE Signal Processing Society (SPS) Image and Multidimensional Signal Processing Technical Committee (IMDSP TC), where he chaired the awards subcommittee (1998–2004). Currently, he is chair of the SPS Multimedia Signal Processing TC, member of the ComSoc Multimedia TC, member of the IEEE SPS Fellow selection committee, and member of the TMM and ICME Steering Committees. He was the founding technical chair for the inaugural NetCod 2005 workshop, special session and panel chair for the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), publicity chair for the Packet Video Workshop 2009, and technical co-chair for the 2009 IEEE International Workshop on Multimedia Signal Processing (MMSP 2009). He is a member of Phi Beta Kappa, Tau Beta Pi, Sigma Xi, and the IEEE Computer, Information Theory, Signal Processing, and Communications societies, and was an active member of the MPEG committee. He is the recipient, with T. Lookabaugh, of the 1993 Signal Processing Society Paper Award; with A. Seghal, of the 2002 ICME Best Paper Award; with Z. Miao, of the 2007 IEEE TRANSACTIONS ON MULTIMEDIA Best Paper Award; and with M. Ponec, S. Sengupta, M. Chen, and J. Li, of the 2009 ICME Best Paper Award.



Bruce Culbertson (Member, IEEE) received the M.A. degree in mathematics from the University of California San Diego, La Jolla, in 1976 and the M.S. degree in computer and information science from Dartmouth College, Hanover, NH, 1981.

He is a Research Manager currently leading a project in continuous 3-D at Hewlett-Packard Laboratories, Palo Alto, CA, where he has worked since 1984. He previously led a project in immersive communication and remote collaboration. He has also worked on projects involving virtual reality, novel cameras and immersive displays, 3-D reconstruction, defect-tolerant computer design, reconfigurable computers, logic synthesis and computer-aided design of digital logic, computer design, and voice and data networks. His current research interests are in computer vision, 3-D, and immersive communication.

Mr. Culbertson has served on the program committee of conferences including the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and the IEEE International Conference on Computer Vision (ICCV).

Ton Kalker (Fellow, IEEE) graduated in mathematics in 1979 and received the Ph.D. degree in mathematics from Rijksuniversiteit Leiden, The Netherlands, in 1986.

He VP of Technology at the Huawei Innovation Center, Santa Clara, CA, leading the research on next generation media technologies. He made significant contributions to the field of media security, in particular digital watermarking, robust media identification, and interoperability of digital rights managements systems. His watermarking technology solution was accepted as the core technology for the proposed DVD copy protection standard and earned him the title of Fellow of the IEEE. His subsequent research focused on robust media identification, where he laid the foundation of the content identification business unit of Philips Electronics, successful in commercializing watermarking and other identification technologies. In his Philips period, he has coauthored 30 patents and 39 patent applications. His interests are in the field of signal and audiovisual processing, media security, biometrics, information theory, and cryptography. He joined Huawei in December 2010. Prior to Huawei, he was with Hewlett-Packard Laboratories, focusing on the problem of noninteroperability of DRM systems. He became one of the three lead architects of the Coral consortium, publishing a standard framework for DRM interoperability in summer 2007. He served for six years as visiting faculty at the University of Eindhoven, Eindhoven, The Netherlands.

Dr. Kalker participates actively in the academic community, through students, publications, keynotes, lectures, membership in program committees and serving as conference chair. He is a cofounder of the IEEE Transactions on Information Forensics. He is the former chair of the associated Technical Committee of Information Forensics and Security.

Mitchell D. Trott (Fellow, IEEE) received the B.S. and M.S. degrees in systems engineering from Case Western Reserve University, Cleveland, OH, in 1987 and 1988, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1992.

He was an Associate Professor in the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, from 1992 until 1998, and Director of Research at ArrayComm, Inc., San Jose, CA, from 1997 through 2002. He is now a Distinguished Technologist at Hewlett-Packard Laboratories, Palo Alto, CA, where he leads the Seamless Collaboration project. His research interests include streaming media systems, multimedia collaboration, multiuser and wireless communication, and information theory.

Susie Wee (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1990, 1991, and 1996, respectively.

She is the Vice President and Chief Technology and Experience Officer of Collaboration at Cisco Systems, San Jose, CA, where she is responsible for driving innovation and user experience in Cisco's collaboration products and services which include unified communications, IP voice and video, telepresence, web conferencing, social collaboration, and mobility and virtualization solutions. Prior to this, she was at Hewlett-Packard Company, Palo Alto, CA, in the roles of founding Vice President of the Experience Software Business and Chief Technology Officer of Client Cloud Services in HP's Personal Systems Group and Lab Director of the HP Labs Mobile and Media Systems Lab. She was the coeditor of the JPSEC standard for the security of JPEG-2000 images and the editor of the JPSEC amendment on File Format Security. While at HP Labs, she was a consulting Assistant Professor at Stanford University, Stanford, CA, where she co-taught a graduate-level course on digital video processing. She has over 50 international publications and over 40 granted patents.

Dr. Wee was an Associate Editor for the IEEE Transactions on Circuits, Systems and Video Technology and for the IEEE Transactions on Image Processing. She received *Technology Review*'s Top 100 Young Innovators award, *ComputerWorld*'s Top 40 Innovators under 40, the INCITs Technical Excellence award, and the *Women In Technology* International Hall of Fame award. She is an IEEE Fellow for her contributions in multimedia technology.