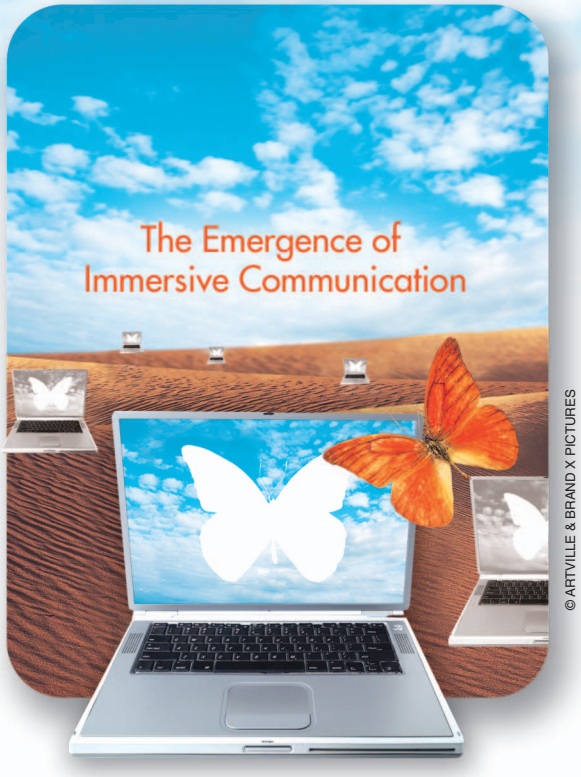
Jeremy R. Cooperstock

Multimodal Telepresence Systems

Supporting demanding collaborative human activities



Significant uses of immersive communication systems include support for group design activities in a distributed context, combining human-human communication and interactivity with synthetic computer-generated content, such as three-dimensional (3-D) models or visualizations. As the demands of distributed human communication and simultaneous interaction with synthetic media continue to expand into new collaborative applications, including surgical planning, simulation training, and entertainment, we see a growing need to improve upon the state of the art of the associated technologies. To this end, we consider the demands on next-generation telepresence systems to support some of the most demanding collaborative human activities. Specifically, we explore the communications requirements for tightly coupled activities in a "shared reality" environment, where distributed users are free to move about their physical spaces, and interact with one another as if physically copresent.

INTRODUCTION

Immersive virtual reality environments (e.g., the CAVE [14]) are unsuitable for group use due to the manner in which the display is rendered for a single viewer (perspective). Issues of fatigue, for example, from shuttered glasses or hand-held orientation sensors, limit long-term usability and the equipment itself often inhibits natural interaction. Videoconferencing suffers from limitations of fidelity and delay and often proves inadequate for supporting group discussion or highly collaborative activity, especially between more than two sites. An additional challenge is the lack of support for gaze awareness. Recent experimental 3-D video teleconferencing technologies [29] overcome these problems by simultaneously rendering multiple views of a remote participant and using a beam splitter to provide reasonable support for eye contact. Higher-end videoconferencing systems (e.g., Cisco Telepresence, HP/Dreamworks Halo, and McGill's Ultra-Videoconferencing technology), which concern themselves solely with the human-human communication problem, have recognized the importance of image size [42] and the presentation of a common visual background to enhance the participants' sense of copresence. However, in the examples above, mobility of participants remains significantly restricted.

These limitations are not solely ones of bandwidth or video quality, but equally of the architecture necessary to support complex, mobile, highly interactive group activity. Research efforts to overcome these limitations necessarily span a number of disciplines, including the acquisition process, low-latency network communication, interaction paradigms, and display technologies. Although video is often emphasized, achieving a

Digital Object Identifier 10.1109/MSP.2010.939040
Date of publication: 17 December 2010

true "telepresence" requires attention to other modalities as well, including audio and the often-neglected haptic channel. Low-fidelity audio of a remote collaborator, delivered from a loudspeaker that is not positioned coherently with the video display, can immediately break the illusion, however weak, of copresence. Similarly, for immersive experiences in which the users are intended to be situated in a pine-covered forest or on a snowy mountaintop, the actual perceptual cues provided to the feet by a linoleum floor surface are hardly convincing. For this reason, it is important to consider directionally sensitive audio acquisition and spatialized rendering [60], as well as haptic synthesis of appropriate ground surface textures [55]. Various successes have been achieved in dealing with the challenges above, individually. However, integrating all of these requirements into a single "shared reality" framework that supports the immersive communication demands for unconstrained group activities in an effective manner, remains an unsolved problem. This article considers the ensemble of technology components necessary to achieve such a goal, illustrating by example the current state of the art.

THE MUSICAL MOTIVATION: FROM DISTRIBUTED MUSIC TO OPERA PERFORMANCE

Despite a traditional focus of immersive communications technologies on high-fidelity videoconferencing and simulation training environments, there are compelling reasons to consider the demands facing distributed musical performance applications. This has been the subject of a considerable number of experiments [32], [8], [48], [58], for the obvious reason that music is one of the most demanding applications available in terms of sensory characteristics and the critical nature of timing. The challenges posed led to an important understanding of timing sensitivity for synchronized activities and ultimately, to notable advances in low-latency network transport protocols [12] for uncompressed media over high bandwidth networks to avoid encoding and decoding delays [58]. More recently, advances in high performance codecs, both for audio and video, have to a large degree obviated the reliance on uncompressed media for time-critical applications. However, other sources of latency persist, such as signal acquisition and display circuitry, buffering, and of course, network transport, always bounded by the speed of light.

When distributed musicians attempt to synchronize with each other, the effect is that the tempo of their playing continuously slows down [8]. Even in a physically copresent scenario, audio propagation delays between musicians can result in tempo drift, which motivates the need for a conductor in symphonic performance, whose visual cues keep the musicians synchronized. Although these problems do not manifest so apparently in a non-musical context, even normal speech interaction can become unnatural and uncomfortable when faced with delays above approximately 200 ms, at which point, turntaking is impeded by collisions of a new speaker with the current speaker [16]. For music, tempo dictates the tolerable latency at which distributed musicians can perform "naturally" and audiences can perceive the performance as synchronized. Obviously, lower latencies support performance of material at a faster tempo. This factor underlies previous efforts to reduce sources of delay in network media streaming, in particular for audio, and motivates ongoing research to reduce the effective latency that each node experiences, possibly by anticipatory encoding of data.

Historically, musical performance has served as one of the "Holy Grail" applications for telepresence systems, given its demands on audio fidelity and latency. Even in the least challenging cases with low network delay and a sufficiently slow musical tempo to permit a successful performance, it is generally recognized that distributed musical performance feels different from the copresent case. Part of this is due to the limited immersion experienced by distributed performers, resulting from a myriad of factors not isolated to audio latency. However, music is often performed with little need for visual fidelity or with support for significant mobility of the performers. It is thus difficult to appreciate the limitations of immersive communication systems in this context.

This observation motivates exploration of similarly demanding areas of human activity for which the interaction between users is more encompassing than synchronized audio and that includes movement beyond a limited workspace. One such candidate application is distributed opera, in which the singers not only engage musically, but theatrically, relying on rich video displays that support eye contact between physical and "virtual" performers to achieve an emotional connection with each other. Moreover, the activity of the performers must be highly synchronized with each other, requiring tightly coupled audio, spatially rendered to provide location cues. Finally, to enrich the presence of the remote performers, as rendered in each location, haptic communication, for example, as conveyed through floor vibrations in response to their footsteps, may be valuable. Distributed opera performance, it can be argued, provides a more meaningful and demanding testbed to explore the limits of immersive communications technology. The discussion of technologies in this article is meant to be fairly general in application. However, our exploration of these technologies was inspired in large part by the communications requirements for a distributed opera performance.

VIDEO

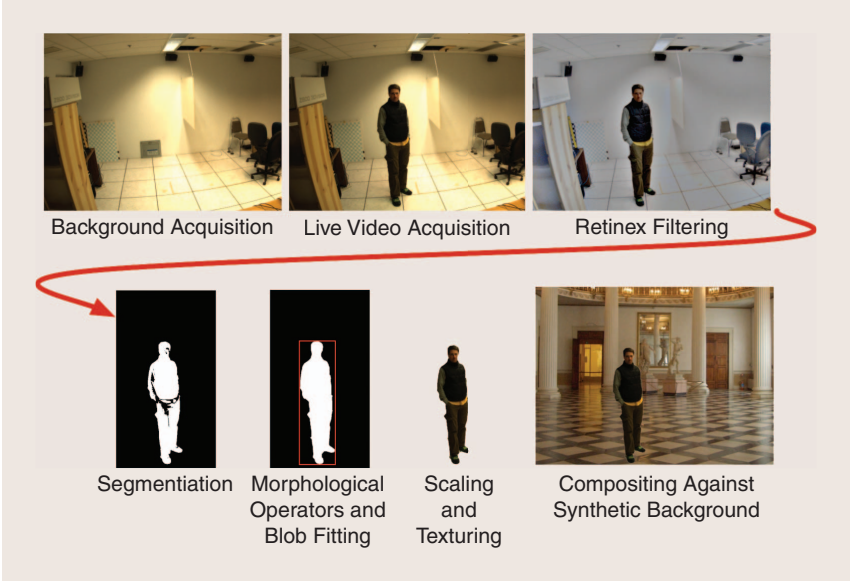
Regardless of application context, live interaction motivates real-time algorithms, as used for encoding, filtering, or synthesis functions. This may impose tradeoffs between quality and speed, as is evident for operations such as video segmentation or image-based rendering, where the best-performing algorithms are often prohibitively expensive. As a concrete example, we describe a representative video processing pipeline to extract a reasonable segmented view of an individual for immersive communications, illustrated in Figure 1, and discuss some of the issues that influence choice of algorithms and hardware.

Latency minimization, discussed in the section "Latency and Synchrony," is critical for interactive applications. Since the

video path begins at a disadvantage in terms of latency compared to other modalities, there is little, if any, allowance for processor-intensive operations through the various stages of this processing pipeline.

ACQUISITION

Any video pipeline begins with acquisition of the raw data, and the choice of camera and lens hardware, equipment placement, frame rate and encoding format. Despite the general availability of camera output in RGB or YUV format, it may be advantageous to acquire single sensor color filter array (CFA) data from the camera in raw Bayer format, leaving the task of color demosaicing to the computer. This is a more significant consideration when dealing with a large number of cameras, where the bandwidth limitations of a particular bus may constitute a bottleneck. A further advantage of working with the original sensor data is that this lends itself better to super-resolution processing, since the aliasing is preserved in the raw data.



[FIG1] **Sample video processing pipeline for segmentation and compositing.**

1) *Single-Camera Approaches*: With respect to immersion, 3-D (stereoscopic) acquisition technology has been discussed at length as a means of achieving greater realism and engagement [45]. Although depth perception is an important aspect of interaction, it remains to be demonstrated that binocular cues (stereopsis and convergence) play a critical role in our interaction with other people. Moreover, our ability to judge depth is usually informed more by monocular cues such as motion parallax, perspective, size, occlusion, shading, and depth-from-motion. Perhaps of greater importance is eye contact and gaze awareness [3], two properties that depend on camera placement relative to the display of the avatar. In typical videoconferencing configurations, the camera is positioned horizontally centered and immediately above the video display. However, the quality of eye contact is dependent on the disparity between the avatar's eyes and the position of the camera, which increase with horizontal displacement from center of the avatar and as the viewer approaches the display. Improved approaches to achieving eye contact include the use of half-silvered mirrors [7], beam splitters (see the review by Grayson and Monk [20]), and video interpolation between multiple cameras, placed at the sides of the display [4]. These are effective in the context of stationary viewers, but unfortunately, both options are limited in the range of viewing angles, and hence, mobility, that can be supported. Moreover, a single video camera is generally inadequate to capture video of multiple individuals at sufficient resolution for high-quality display, in particular when they are not clustered together.

2) *Camera Arrays*: Addressing the challenges of acquisition, either for rendering in 3-D or from an arbitrary desired viewpoint, might well entail camera arrays that can sample the environment from a wide range of positions and poses. Since switching between views abruptly would result in a visual discontinuity, we can instead synthesize intermediate virtual camera views dynamically between these real cameras, using an image-based modeling and rendering method [49]. Such methods may employ an explicit 3-D geometry description of the scene or person being rendered or use correspondences between views to establish the geometry implicitly. Obtaining the geometry is often computationally expensive [21], but can be simplified using techniques such as structured light [15], [45] and the image-based visual hull [36]. A promising approach for real-time view interpolation involves the use of a GPU-based implementation of a plane sweep depth estimation [11]. This involves projecting the images from a set of cameras onto a set of parallel planes, then using some photoconsistency metric to select the "best" plane for each pixel, implicitly determining the actual 3-D scene description that gave rise to the observations. In the case of occlusions, some mechanism may be required to select which cameras are included in the set. The smoothness of the resulting depth map [62] is often a critical factor in the quality of the reconstructed image. A balance must be struck between reproducing geometric detail and avoiding the introduction of visual artifacts due to noise in these depth estimates. Challenges to obtaining an accurate depth map include the presence of regions of uniform texture, specularities, or very fine detail, all of which pose problems for photoconsistency measures. One possible solution is to integrate the depth information obtained from a time-of-flight camera, which is insensitive to scene texture [63]. This allows for synthesis of an intermediate viewpoint with a significantly

(a)

(b)

[FIG2] Performers displayed on different display technologies (a) Bang and Olufsen 103" plasma and (b) DNP Holo screen.

reduced need for smoothing of the depth values to reduce noise. Generation of a depth map, using either passive or active techniques, can also prove useful for segmentation purposes, as discussed in the section "Image Filtering and Segmentation."

3) *Illumination*: Illumination is often an important consideration as well. Adequate lighting is required to integrate each frame in the camera sensor, but such lighting can often detract from the immersive experience. Fortunately, in typical performance situations, stage lighting does offer sufficient illumination, but this cannot be assumed in general. Gross et al. [21] describe a video acquisition process using switched banks of white LEDs that illuminate in bursts, timed to coincide with periods during which the participants' shutter glasses are opaque. Unfortunately, this requires the use of glasses by all participants, and more problematic, only captures their video when the glasses are opaque, thus, with their eyes not visible. Obviously, this is unacceptable in the context of immersive communications.

IMAGE FILTERING AND SEGMENTATION

It is often desirable to compensate for uneven or otherwise poor illumination conditions, using homomorphic filtering, such as the popular multiscale retinex filter [34]. This provides color consistency of the output in a manner that better matches the characteristics of the human visual system.

Although simply displaying the filtered video output would be suitable for videoconferencing, this does not achieve the intended immersive effect, and thus detracts from the engagement of a performance or simulation application. To enhance the perception of copresence, foreground video of the performers or participants must be segmented from the background. This facilitates video compositing on a synthetic background or rendering a convincingly realistic avatar in the scene on a transparent display surface such as the Holoscreen, as illustrated in Figure 2.

Foreground-background segmentation has long been an important problem in computer vision, with applications spanning tracking to augmented reality [56]. As a result, a wide variety of algorithms exist, based on color matching, wavelet transforms, Gaussian mixture models [51], and motion models, among others. A number of these are computationally expensive or are unable to cope with fast variation of the background, and must therefore be rejected for real-time applications. Further challenges to such algorithms include illumination variation, dynamic background activity, occlusion, and shadows. Robust segmentation often requires additional equipment, such as a structured light source [40], foreground or background infrared illumination [61], or range measurements, e.g., from a time-of-flight camera [13] to obtain a clear boundary of the foreground object, in particular where the background is dynamic.

In the examples shown throughout this article, we adopted an optimized codebook algorithm [30], as implemented in OpenCV. During training, this approach generates a set of codewords for each pixel, representing simple statistics such as the frequency of a particular intensity value. Subsequently, the distance of a pixel to the cluster of codewords is used to determine whether it belongs to the background. This proved to be faster than the mixture-of-Gaussian or kernel methods, yielded generally superior results, and was also reasonably robust to dynamic backgrounds and illumination variation.

Since the segmentation results are likely to exhibit some noise, additional morphological image processing operators, typically involving dilation and erosion, are employed to obtain a smooth binary mask, free of holes, representing the pixel locations of the foreground individual.

BLOB DETECTION AND TRACKING

The next step in the pipeline is blob detection. This is necessary both for the following step of video blending, but also, in anticipation of the eventual need for knowledge of foreground object

motion in dynamic video mosaicing operation or to facilitate acquisition from pan-tilt cameras to support effective rendering of moving individuals, displayed at reasonable resolution. In this case, the pan-tilt values are combined with the position of the blob in the video frame to determine the actual position.

Blob detection and tracking can be trivial operations, in particular following the steps described above, since, in theory, these only require only a scan through each frame for pixels that have been retained by the segmentation operation. However, under more realistic conditions of additional scene content and multiple dynamic participants that may occlude one another, these steps can become more complicated. Following initialization on one or more desired blobs, Kalman filter methods, along with various generalizations such as the particle filter, can be applied to cope with these challenges. Latency considerations must also be kept in mind; accuracy of tracking, which may be improved by using a larger number of particles, generally entails greater computational cost.

SCALING AND BLENDING

The final step in the rendering pipeline may involve blending or "compositing" the textured video content into a virtual scene. For example, remote performers may appear in front of a synthetic background, visible to the audience. Determining the appropriate mapping between pixel positions in the camera frame of reference and the corresponding rendered position of the individual in the virtual scene must take into account issues of camera position, lens focal length, dimensions of the physical environment, and that of the virtual background.

DISPLAY

Various video display options exist for immersive communications. Common technologies include monitors (cathode ray tube (CRT), liquid crystal display (LCD), and plasma), projection [typically LCD or digital light processing (DLP)], and head-mounted displays, but there also exist more esoteric options such as retinal projection [53] and several "cheats" such as Pepper's Ghost [57] and "Holo screens" that are optically transparent but support the projection of a clear image on the surface, thus rendering a two-dimensional (2-D) hologram-like effect [Figure 2(b)].

Both monitors and projector arrays can be employed to produce 3-D effects, for example, using polarized glasses, shutter glasses, and head-mounted displays. Although the use of glasses has been the norm for virtual reality experiences, this is less acceptable for scenarios involving human-human interaction, as glasses restrict visibility of the wearer's eyes. Autostereoscopic technologies such as multiview lenticular displays and parallax barrier displays avoid this problem by presenting an image that is dependent on view angle, often supporting multiple users simultaneously, each with an independent view. Some technologies, such as head-mounted displays and retinal projection, are suitable only for a single viewer at a time, others are better suited for simultaneous viewing by multiple users, and some are capable of supporting stereo viewing without requiring glasses, such as autostereoscopic and swept-surface volumetric displays. However, constraints on viewing angle and distance often limit viewer mobility.

The choice of display technology can impact strongly on the quality of telepresence. Screen bezels are on obvious distraction that may break the sense of immersion in a virtualized environment. So too are display brightness, which limits allowable ambient illumination and field of view and constrains the viewing angle, beyond which the remote participant, or avatar, effectively disappears. In our application domain of a distributed opera, the chosen technology should provide a perceptually satisfactory view both to the audience members and to other on-stage performers, who may be positioned at arbitrary locations relative to the display. This would limit suitable options to those permitting correct views from a full 360°, such as fog displays [44], 3-D swept-surface volumetric displays, such as those by Actuality, Felix 3-D, and Genex, or the slice-stacked rear-projection Lightspace technology [52], and computer-generated holography [50]. The impressive volumetric display prototype demonstrated by Jones et al. [29] suggests promise in this regard, especially once scaling and mobility issues are addressed to provide an equivalent display to represent a full-size human, who can move freely about the stage. Until then, however, rendering in such a manner that remote performers appear correctly both to other performers nearby on-stage and to audience member at a distance will remain a daunting challenge.

LATENCY AND SYNCHRONY

Video often represents the greatest source of latency in distributed applications, due to its commensurately higher bandwidth than the other modalities. This often necessitates compression prior to transmission, and the integration time associated with a single frame. In our efforts to reduce these latencies, we have investigated various techniques, including uncompressed video transmission over network lightpaths and transfer of video data in subframe chunks between the video interface and network interface devices before the entire frame has been received. Working with progressive scan 720p60 (SMPTE 296M) high-definition video, our Ultra-Videoconferencing software achieves total end-to-end delay, which includes camera acquisition, (local) network transmission, and display on a monitor, of 52 ms, which represents an additional latency of only 22 ms beyond the minimum camera-to-monitor direct connection time. These figures are measured by video recording of an electronic flash that is viewed both by the high-speed recording camera and the high-definition camera. The flash thus appears both "live" and on the video monitor connected to the receiving computer.

For most interactive activities, video latencies on this order are considered acceptable and often imperceptible. Even high-end commercial videoconferencing systems exhibit network delays of 150 ms, exclusive of codec processing time, considering this acceptable for conversational interaction (see http://www.cisco.com/en/US/docs/solutions/Enterprise/Video/tpqos.html). However, for telerobotic laparoscopic surgical tasks employing video feedback only, studies have demonstrated

significant effects on task completion time and error rate as latencies increase from zero delay to 175 ms and above [2].

AUDIO

Despite a predominance of research attention to the video medium as a form of immersive communication, we should not overlook the importance of other modalities to the perception of immersion. The most obvious, in particular from the context of the musical application described above, is audio. In this regard, film soundtracks are routinely used to evoke intended emotional responses, but even in isolation from other media, music has been demonstrated to have significant effects, including the experience of "chills" [39]. Beyond the factors of sampling frequency and resolution, we consider the ability to localize the spatial source of a sound as a critical factor for immersive communications. The remainder of this section addresses such audio-specific issues.

ENCODING

Well-known standards for encoding of surround audio are Dolby Digital AC-3 and Digital Theater System (DTS) formats, as used for traditional 5.1-channel DVD audio storage. Both these and several compressed formats, including Moving Picture Experts Group (MPEG) Surround and Ogg Vorbis, have been demonstrated for use in network streaming applications, thereby providing audio location awareness capabilities for communication. However, these formats are fixed to a specific loudspeaker topology. Ambisonics encoding additionally includes specification of the vertical dimension, thus allowing for reproduction of the full-sphere sound field, as described below. These encoding formats do not permit sound source separation or dynamic adjustment of the sound locations at the receiver end.

MPEG-4 Audio Binary Format for Scenes (BIFS), based on VRML, and X3D, the successor to VRML, both provide a 3-D scene description standard that includes audio. Similarly, the upcoming MPEG standard, Spatial Audio Object Coding (SAOC) [26], was developed as a parametric coding technology for perceptually relevant spatial properties. SAOC is intended to provide high-quality surround audio using similar bandwidth as previously required for traditional two-channel stereo [23]. Importantly, since these encoding methods are not tied to any particular output configuration, they allow the decoding side to control aspects of the spatial audio rendering.

TRANSMISSION

Audio data is transported over networks using a communication protocol, often the standards-based Real-Time Transport Protocol (RTP). To minimize latency and connection overhead, most such audio transport protocols employ the lower-level User Datagram Protocol (UDP). However, there are some exceptions such as Adobe's Real-Time Messaging Protocol (RTMP), which uses the connection-oriented Transmission Control Protocol (TCP) instead. This offers the benefit of congestion avoidance and is better suited for bypassing firewalls, but generally suffers higher latency than UDP-based protocols.

Since delay or loss of a packet of audio data can be disruptive to the listener experience, various sender-based repair strategies, including retransmission protocols and forward error correction (FEC), have been proposed. Retransmissions can be problematic because of the additional round-trip network delay that is imposed, whereas FEC adds redundancy to the stream although at the expense of greater bandwidth requirements. Alternatively, loss concealment techniques have been investigated to mask the effects of missing audio data at the receiver side. These range from the simple substitution of silence or white noise for the missing content to perceptually superior methods involving interpolation and regeneration. For speech audio, loss concealment has been found to be effective for up to 15% packet loss when packets contain between 4 and 40 ms of audio [41]. Corresponding measurements for music are highly dependent on genre.

REPRODUCTION

Audio rendering or reproduction for virtual environments draws on a rich background from spatial auditory displays and psychoacoustics [5]. Synthetic audio displays can be computed by a variety of physics-based software that include modeling of distance delay, room response, Doppler effects, and reverberation [38], [59].

The audio display itself can be rendered through headphones or loudspeakers. The former provide greater immunity to ambient sound and the acoustic properties of the reproduction environment. Headphones can deliver a personalized output that is appropriate to each listener's position and orientation, assuming these can be tracked reliably. Moreover, headphones isolate the desired signal to each ear, thereby avoiding crosstalk. To achieve the desired sound imaging characteristics, binaural displays apply frequency-based filtering according to a model of the user's head response transfer function (HRTF). These synthesize the effects that physical sound source location would have on the experience of the listener. Stereo or quadraphonic loudspeaker configurations with crosstalk cancellation can achieve similar spatial audio reproduction, albeit with constraints on allowable motion of the listener.

For general performance applications such as the distributed opera scenario described above, use of headphones is impractical and the loudspeaker alternative must be considered. Here, the challenge is to create a compelling immersive experience for audience members, while simultaneously supporting on-stage communication between performers. Solutions typically employ stage monitors and sound reinforcement techniques developed by the audio engineering community. Multichannel loudspeaker configurations can provide a static frame of reference for the rendered sound sources, implying that directionality is preserved regardless of user orientation. However, a tradeoff must be considered between the quality of localization and the width of the "sweet spot," outside of which the sound signals are no longer properly focused, thus degrading the quality of the resulting sound field.

Conventional loudspeaker configurations are on a single plane. Such systems are thus limited to reproducing the 2-D

sound field, parallel to the ground, ignoring the vertical dimension. For most screen- or stage-based applications, activity takes place approximately at eye level, so localization in the horizontal plane is the most important. However, human audio perception is sensitive to height cues and these can enhance the degree of immersion when reproduced by loudspeakers placed at two or more planes. In this case, the rendering system must know how to translate the audio data into an appropriate mix of signals to the loudspeakers to generate the intended 3-D spatial audio sound field.

The simplest form of spatial audio rendering is amplitude panning, as used in conventional stereophonic configurations. Perceived direction, or panning angle, of the sound source is determined by the relative gain of each loudspeaker. This approach was generalized to arbitrary configurations with vector base amplitude panning (VBAP) [43], which renders each virtual sound source as a linear combination of three loudspeakers (or two in the case of 2-D reproduction). A related technique, time panning, can also be employed, as perceived direction of a virtual sound source is influenced by the interaural arrival times of the signal. However, this is frequency dependent and thus, less often used. Amplitude panning is implemented in most software toolkits for spatial audio rendering, including Creative Lab's OpenAL and Microsoft's DirectSound. Improved localization cues can be provided through the addition of early reflections [25].

Ambisonics [18], based on Huygens' principle, offers the advantage over conventional rendering techniques of treating sound sources as isotropic. Although limited in the size of the listening area for correct sound field reproduction, it nevertheless preserves the characteristics of the rendered sound field more stably with listener position than do most other surround-sound techniques. In its basic form, the sound field is encoded in a four-channel representation consisting of a nondirectional monophonic pressure signal W, captured from an omnidirectional microphone, and three directional gradient signals, front/back X, left/right Y, and up/down Z, captured by figure-eight (bidirectional) microphones. Rendering of a full sphere requires a minimum of six loudspeakers. Higher-order ambisonics incorporate additional encoding channels for greater localization ability and improved performance in large environments.

Wave field synthesis (WFS) [6], differs from ambisonics in that it allows for sound sources to be localized independently of listening position, thus providing an effective experience over almost the entire area of rendering rather than being constrained to the "sweet spot." WFS, also based on Huygens' principle, simulates the wavefront of primary audio sources using the superposition of spherical waves emitted by a large number of loudspeakers. These are arranged in a circular array, with minimal separation to avoid audible spatial aliasing. With present technology, WFS is only practical in a single plane.

LATENCY AND SYNCHRONIZATION

Minimization of audio latency can be critical for immersive communication in simulation and distributed performance applications. For distributed communication involving speech, spontaneous interruption becomes problematic with latencies above 100 ms and speech ceases to feel "natural" above 200 ms. For more tightly coupled interactive activities, such as musical performance, the timing requirements are often more demanding, with ensemble piano performance on fast movements becoming difficult with latencies as low as 10 ms [9]. As latency increases, how tempo naturally decreases in network musical performance unless the performers ignore their remote counterpart(s). These factors impose constraints on allowance for processing times and network transport configurations. For the latter, a sufficient buffer is required to smooth out the effects of variations between packet arrival times, typically associated with network jitter.

In contrast to video, audio can be transferred end-to-end with a total latency as low as 5 ms, provided that the audio interface allows for specification of very small buffers. In practice, a value of 10 ms is necessary to ensure reasonable stability. Regardless, this is significantly less than the minimum video delay typically achievable.

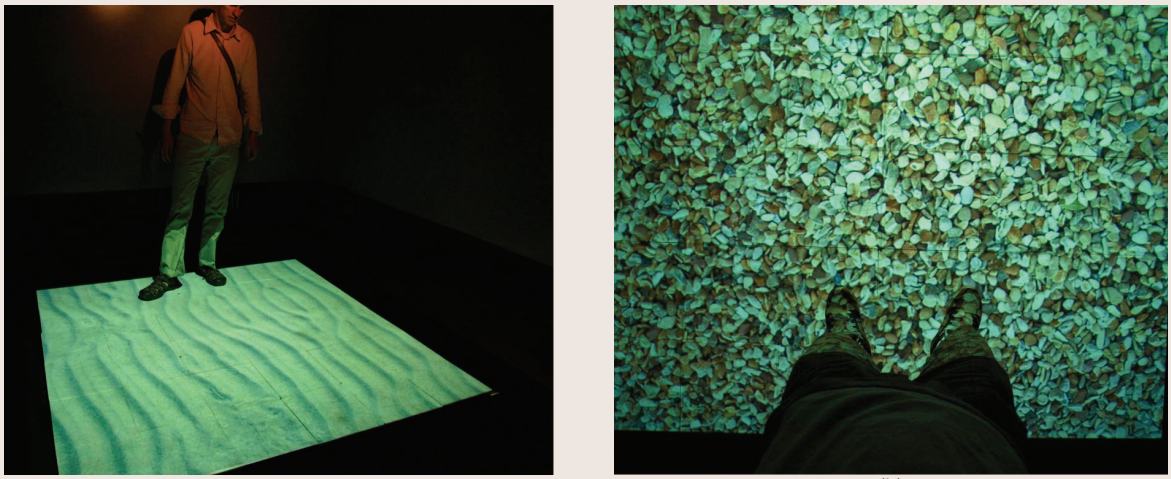
Immersive communications involving audio in conjunction with graphics or video must also consider synchronization between streams to avoid disrupting the immersive experience. Since audio lags video in nature, the relative misordering of these cues in a mediated communication context may be problematic with respect to immersion. For example, lip-sync problems become apparent when audio lags video by more than approximately 125 ms but when audio leads video, this figure drops to approximately 45 ms [27]. As such, it may prove advantageous to add delay to the audio signal so that it does not lead the video [46].

HAPTICS

Haptics, whether involving kinesthetic (force) or vibrotactile feedback, can play a significant role in enhancing the quality of immersive experience. To date, this has been most evident in teleoperation tasks, in which forces sensed by a remote slave are relayed back to the human operator of the master controller. The relevance of the haptic modality also extends to other realms of immersive communication, for example, simulating the ground surface properties of a simulated environment or providing awareness of distributed group activity. The remainder of this section discusses haptics in these contexts and the associated challenges.

ENCODING AND TRANSPORT

Vibrotactile haptic data, often obtained from force sensors or accelerometers, may be encoded and transported as if it were a low-frequency audio signal. However, unlike audio, haptic communication often requires a 1 kHz control loop, which implies a high rate of transmission of small network packets. To alleviate these demands, haptic codecs often employ bit rate reduction, taking advantage of psychophysical knowledge. Since haptic perception, like that of our other senses, follows Weber-Fechner's law of logarithmic response to the induced stimuli, there is no need to transmit changes in the signal



[FIG3] **Sample terrains rendered with haptic texture are shown in parts (a) and (b). Figure courtesy of Y. Visell.**

unless they exceed a perceivable threshold. This principle leads to "deadband" coding, for which packet rate reductions of 90% have been reported, while preserving transparency of telerobotic systems [24]. Data transport itself can be accomplished with the same protocols used for other media, including the popular RTP [31].

DISPLAY

To date, most examples of haptic display involve either whole body vibration or local stimulation of the fingers and hands. Popular user interface devices in the latter category include the Phantom (SensAble), TouchSense (Immersion), Rumble Pak (Nintendo), and the DualShock gamepad (Sony). Aside from video games [17], applications are extending increasingly to virtual environments, theme park and home theatre entertainment, medical, dental, and avionics simulation. More recently, investigation of the possibilities of haptic feedback for immersive experiences has expanded to body-worn gaming vests (a noteworthy example is the University of Pennsylvania's Tactile Gaming Vest. See http://spectrum.ieee.org/automaton/robotics/robotics-software/tactile-gaming-vest-punches-and-slices for further details) and full-body experience of realistic wind flow [33].

The relevant haptic sensations necessary to augment one's experience of a synthetic or remote environment can also be conveyed effectively through the feet, which are often in contact with the ground surface. In the domain of distributed performance, some experiments have been conducted using vibration delivered through a motion platform. These efforts explored the delivery of haptics to provide audience members with an improved perception of presence, either for music [58] or dance. Haptics can convey information in iconic form, for example, as alerts or warnings, or serve as a means of explicit communication of emotionally salient information such as a loved one's heartbeat or the kicking of a fetus in the mother's womb, which we have reproduced through a motion actuator. Successful identification of ground material is, unsurprisingly, most strongly correlated with the haptic modality [19]. In copresent interaction, haptic information not only informs users of ground material properties, but may provide subtle cues that guide attention, for example, regarding the activity of other individuals through footstep vibrations.

Our investigation of the role of haptics in a multimodal immersive environment culminated in our prototypes of various terrains, including a snow-covered landscape [35] and a virtual ice pond [54]. These simulations are particularly engaging because they provide feedback through all the modalities of sight, sound, and touch. The action of stepping on the virtual materials produces similar visual deformations, auditory responses, and textural vibration effects as physical snow compaction or ice fracture. The importance of the haptic modality can be appreciated with reference to the surfaces depicted in Figure 3.

LATENCY AND SYNCHRONY

Similar issues apply to visuo-haptic synchrony as they do to audiovisual modalities. Haptic and proprioceptive asynchrony with other modalities are equally can break the "illusion" of immersion, or worse, give rise to cybersickness in head-coupled virtual reality cybersickness [37]. In virtual reality and teleoperation tasks, the haptic channel provides a feedback loop, for which communication latency may impact on stability, thereby degrading task performance. When distributed users interact, variable network delays can result in disagreement in the position and orientation between the multiple representations of a shared object. This impedes the users' ability to collaborate on a cooperative task. In such cases, coherency can be achieved by virtual coupling strategies [47].

INTERACTION CHALLENGES

Manipulation of nonhaptic input devices, e.g., keyboard and mouse, only involves the application of forces by the user. Although feedback latency can be disruptive, this will not

physically degrade the interaction [1]. Similarly, purely vibrotactile displays are generally free of stability problems, since these tend to be driven in an open loop. However, interaction between a user and a haptic system providing kinesthetic feedback involves bilateral control. This can lead to instability, stemming from physical limitations of the underlying hardware, user compensation, various control systems issues, and communication delays including processing time between sensor measurements and user responses at the display [28]. Such instability degrades the realism of the immersive experience and can potentially be dangerous. There is a tradeoff between stability and performance, as dictated by the dynamic range of impedance that can be rendered passively, or Z-width, of the device [10]. Similarly, stability considerations must be weighed against transparency, which relates to the degree to which the user is unaware of the physical characteristics of the haptic rendering mechanism. In this context, transparency is a major determinant of user immersion. Stability can be obtained by modeling the system and adapting controller parameters based on sensory input related to the human movement or by a number of energy transfer analyses based on passivity theory [22].

CONCLUSIONS

Having surveyed a range of video, audio, and haptics technologies, along with the challenges associated with their use in immersive communication, we now return to the question of what is possible today. Despite the numerous challenges posed by delay and signal fidelity, the feasibility of high-fidelity interaction for simulation training applications and wide-scale distributed musical performance, including opera, over the Internet, appears promising. Notwithstanding past and recent success in these areas, we are concerned with the objective of achieving a result that is artistically and aesthetically successful, and ideally, one that does not require accommodation by the artists to the technology. To that end, we face continued challenges regarding the perception of copresence posed by network latency and limitations in acquisition and display technology.

The speed of light remains the ultimate limiting factor in communications. This suggests that the challenges of latency in distributed interaction for highly time-sensitive applications is expected to persist, resulting in a lack of transparency or naturalness of the experience. However, for certain tasks, latency compensation may prove reasonably effective. Specifically, anticipatory signal synthesis can "predict" future audio, video, and haptic signals based on current observations and prior knowledge of the task. As a first step in this direction, we are exploring the possibility of predicting the baton movements and gestures of an orchestral conductor to render an avatar representation that leads the actual conductor by a fraction of a second.

In this regard, the ongoing challenges we have highlighted include the multimodal display of a remote individual that can move seamlessly, just as an opera performer enjoys full mobility about the stage, in such a manner that the rendering appears perceptually correct from a wide range of viewing angles; and the end-to-end exchange of multiple communication modalities with sufficiently low latency and high fidelity as to permit distributed interaction with an ease that approaches that of physical copresence. Achieving these objectives may require advances in computer-generated holography, the development of novel electro-optical display technologies, and more powerful prediction techniques to cope with latency. Given the stringent technical and aesthetic requirements of multiperson artistic performance, satisfying the associated demands in the context of distributed opera ensures that the technologies developed for this purpose can be applied directly to a wide range of other telepresence applications.

ACKNOWLEDGMENTS

The work described in this article owes much of its inspiration to Niels W. Lund, who conceived the World Opera Project as a new challenge for distributed performance research. Various hardware and software components described in this article were the work of Mitchel Benovoy, Cristina Darolti, Donald Dansereau, Alvin Law, Stéphane Pelletier, Stephen Spackman, Haijian Sun, Yon Visell, and Guangyu Wang. The author is grateful to reviewers for their helpful suggestions regarding the content, and to Nicolas Bouillot, Guillaume Millet, Martin Otis, and Mike Wozniewski for their valuable comments. These efforts were supported by generous funding from the Canarie Inc. ANAST program, the Natural Sciences and Engineering Research Council of Canada, the Norwegian Research Council, les Fonds de recherche sur la nature et les technologies, and the Ministère du Développement, économique, Innovation et Exportation (Quebec).

AUTHOR

Jeremy R. Cooperstock (jer@cim.mcgill.ca) directs McGill University's Shared Reality Lab, which focuses on computer mediation to facilitate high-fidelity human communication and the synthesis of perceptually engaging, multimodal, immersive environments. His accomplishments include the world's first Internet streaming demonstrations of Dolby Digital 5.1, uncompressed 12-channel 96kHz/24-b, multiple simultaneous streams of uncompressed high-definition video, and a simulation environment that renders graphic, audio, and vibrotactile effects in response to footsteps. His work on the Ultra-Videoconferencing system won the Most Innovative Use of New Technology Award from ACM/IEEE Supercomputing and an Audio Engineering Society Distinction Award.

REFERENCES

[1] R. Adams and B. Hannaford, "Stable haptic interaction with virtual environments," *IEEE Trans. Robot. Automat.*, vol. 15, no. 3, pp. 465–744, 1999.

[2] M. Anvari, T. Broderick, H. Stein, T. Chapman, M. Ghodoussi, D. W. Birch, C. Mckinley, P. Trudeau, S. Dutta, and C. H. Goldsmith, "The impact of latency on surgical precision and task completion during robotic-assisted remote telepresence surgery," *Comput. Aided Surgery*, vol. 10, no. 2, pp. 93–99, 2005.

[3] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge, U.K.: Cambridge Univ. Press, 1976.

[4] H. H. Baker, D. Tanguay, I. Sobel, D. Gelb, M. E. Goss, W. B. Culbertson, and T. Malzbender, "The coliseum immersive teleconferencing system," in *Proc. Int. Workshop Immersive Telepresence*, Dec. 2002.

[5] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. San Diego, CA: Academic, 1994.

[6] A. J. Berkhout, "A holographic approach to acoustic control," *J. Audio Eng. Soc.*, vol. 36, no. 12, pp. 977–995, 1988.

[7] W. Buxton and T. Moran, "Europarc's integrated interactive intermedia facility (IIIF): Early experience," in *Proc. IFIP WG 8.4 Conf. Multi-User Interfaces and Applications*, S. Gibbs and A. Verrijn-Stuart, Eds. Amsterdam: Elsevier, 1990, pp. 11–34.

[8] C. Chafe, M. Gurevich, G. Leslie, and S. Tyan, "Effect of time delay on ensemble accuracy," in *Proc. Int. Symp. Musical Acoustics*, 2004, pp. 207–211.

[9] E. Chew, A. A. Sawchuk, R. Zimmerman, V. Stoyanova, I. Tosheff, C. Kyriakakis, C. Papadopoulos, A. R. J. François, and A. Volk, "Distributed immersive performance," in *Proc. Annu. Nat. Association of the Schools of Music (NASM)*, San Diego, CA, 2004, pp. 85–93.

[10] J. E. Colgate and J. M. Brown, "Factors affecting the z-width of a haptic display," in *Proc. IEEE Int. Conf. Robotics and Automation*, San Diego, CA, Étas-Unis, May 1994, vol. 4, pp. 3205–3210.

[11] R. Collins, "A space-sweep approach to true multi-image matching," *IEEE Comput. Vis. Pattern Recognit.*, pp. 358–363, 1996.

[12] J. R. Cooperstock and S. P. Spackman, "The recording studio that spanned a continent," in *Proc. IEEE Int. Conf. Web Delivering of Music (WEDELMUSIC)*, 2001, pp. 161–167.

[13] R. Crabb, C. Tracey, A. Puranik, and J. Davis, "Real-time foreground segmentation via range and color imaging," in *Computer Vision and Pattern Recognition Workshop,* Anchorage, AK, 2008, pp. 1–5.

[14] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti, "Surround-screen projection-based virtual reality: The design and implementation of the CAVE," in *Proc. Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH)*. New York: ACM, 1993, pp. 135–142.

[15] J. W. Davis and A. F. Bobick, "SIDEshow: A silhouette-based interactive dual-screen environment," Massachusetts Inst. Technol. Media Lab., Cambridge, MA, Tech. Rep. 457, 1998.

[16] A. Dix, "Network-based interaction," in *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, A. Sears and J. A. Jacko, Eds. Hillsdale, NJ: L. Erlbaum Associates Inc., 2002, pp. 331–357.

[17] M. Elgan, "Haptics: The feel-good technology of the year," *Computerworld*, July 2009. [Online]. Available: http://www.computerworld.com/s/article/9135897/Haptics_The_feel_good_technology_of_the_year

[18] M. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[19] B. Giordano, S. McAdams, Y. Visell, J. Cooperstock, H.-Y. Yao, and V. Hayward, "Non-visual identification of walking grounds," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, p. 3412, May 2008.

[20] D. M. Grayson and A. F. Monk, "Are you looking at me? Eye contact and desktop video conferencing," *ACM Trans. Comput.-Hum. Interactions*, vol. 10, no. 3, pp. 221–243, 2003.

[21] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. V. Moere, and O. Staadt, "blue-c: A spatially immersive display and 3D video portal for telepresence," in *Proc. Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH)*. New York: ACM, 2003, pp. 819–827.

[22] B. Hannaford and J.-H. Ryu, "Time domain passivity control of haptic interfaces," in *Proc. IEEE Int. Conf. Robotics and Automation,* Seoul, 2001, vol. 2, pp. 1863–1869.

[23] J. Herre and S. Disch, "New concepts in parametric coding of spatial audio: From sac to saoc," in *Proc. IEEE Int. Conf. Multimedia and Expo*, July 2007, pp. 1894–1897.

[24] 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 Processing*, vol. 56, no. 2, pp. 588–597, Feb. 2008.

[25] J. Huopaniemi, "Virtual acoustics and 3-D sound in multimedia signal processing," Ph.D. dissertation*,* Dept. Elect. and Commun. Eng., Helsinki Univ. Technology, Helsinki, Finland, 1999.

[26] C. Ishii, Y. Nishitani, and H. Hashimoto, "Lyapunov function-based bilateral control for teleoperation system with time varying delay," in *Int. Conf. Industrial Technol.*, Mar. 2010, pp 1694–1699.

[27] International Telecommunication Union/ITU Radiocommunication Sector, "Relative timing of sound and vision for broadcasting," Geneva, Switzerland, Tech. Rep. BT.1359-1, Jan. 1998.

[28] C. Ishii, Y. Nishitani, and H. Hashimoto, "Lyapunov function based bilateral control for teleoperation system with time varying delay," in *Proc. IEEE Int. Conf. Industrial Technology*, Mar. 2010, pp. 1694–1699.

[29] 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," in *Proc. Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH)*. New York: ACM, 2009, pp. 1–8.

[30] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, "Background modeling and subtraction by codebook construction," in *Proc. Int. Conf. Image Processing*, Oct. 2004, vol. 5, pp. 3061–3064.

[31] H. King, B. Hannaford, J. Kammerl, and E. Steinbach, "Establishing multimodal telepresence sessions using the session initiation protocol (SIP) and advanced haptic codecs," in *Proc. IEEE Haptics Symp.*, Mar. 2010, pp. 321–325.

[32] D. Konstantas, Y. Orlarey, O. Carbonel, and S. Gibbs, "The distributed musical rehearsal environment," *Multimedia*, vol. 6, no. 3, pp. 54–64, July—Sept. 1999.

[33] S. Kulkarni, C. Fisher, E. Pardyjak, M. Minor, and J. Hollerbach, "Wind display device for locomotion interface in a virtual environment," *World Haptics Conf.*, pp. 184–189, 2009.

[34] E. H. Land, "An alternative technique for the computation of the designator in the retinex theory of color vision," *Proc. Nat. Acad. Sci. USA*, vol. 83, no. 10, pp. 3078–3080, May 1986.

[35] A. W. Law, B. V. Peck, Y. Visell, P. G. Kry, and J. R. Cooperstock, "A multimodal floor-space for experiencing material deformation underfoot in virtual reality," in *Proc. IEEE Int. Workshop Haptic Audio Visual Environments and Games*, Oct. 2008, pp. 126–131.

[36] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," in *Proc. 27th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'00)*. New York: ACM Press/Addison-Wesley, 2000, pp. 369–374.

[37] M. E. McCauley and T. J. Sharkey, "Cybersickness: Perception of self-motion in virtual environments," *Presence: Teleoper. Virtual Environ.*, vol. 1, no. 3, pp. 311–318, 1992.

[38] M. Naef, O. Staadt, and M. Gross, "Spatialized audio rendering for immersive virtual environments," in *Proc. ACM Symp. Virtual Reality Software and Technology (VRST'02)*. New York: ACM, 2002, pp. 65–72.

[39] J. Panksepp, "The emotional sources of "chills" induced by music," *Music Perception*, vol. 13, no. 2, pp. 171–207, 1995.

[40] J. Park, C. Kim, J. Na, J. Yi, and M. Turk, "Using structured light for efficient depth edge detection," *Image Vis. Comput.*, vol. 26, no. 11, pp. 1449–1465, 2008.

[41] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet-loss recovery techniques for streaming audio," *IEEE Network Mag.*, vol. 12, pp. 40–48, Sept./Oct. 1998.

[42] A. Prussog, L. Mühlbach, and M. Böcker, "Telepresence in videocommunications," in *Proc. 38th Annu. Meeting Human Factors and Ergonomics Society*, 1994, pp. 180–184.

[43] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[44] I. Rakkolainen and K. Palovuori, "Interactive digital fogscreen," in *Proc. 3rd Nordic Conf. Human-Computer Interaction (NordiCHI '04)*. New York: ACM, 2004, pp. 459–460.

[45] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, "The office of the future: A unified approach to image-based modeling and spatially immersive displays," in *Proc. Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH)*. New York: ACM, 1998, pp. 179–188.

[46] B. Reeves and D. Voelker, "Effects of audio-video asynchrony on viewer's memory, evaluation of content and detection ability," Stanford Univ., Stanford, CA, Tech. Rep., Oct. 1993.

[47] G. Sankaranarayanan and B. Hannaford, "Experimental internet haptic collaboration using virtual coupling schemes," in *Proc. IEEE Haptics Symp.,* Piscataway, NJ, 2008, pp. 259–266.

[48] A. A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis, "From remote media immersion to distributed immersive performance," in *Proc. 2003 ACM SIGMM Workshop on Experiential Telepresence (ETP '03)*. New York: ACM Press, pp. 110–120.

[49] H. Y. Shum, S. C. Chan, and S. B. Kang, *Image-Based Rendering*. New York: Springer-Verlag, 2007.

[50] C. Slinger, C. Cameron, and M. Stanley, "Computer-generated holography as a generic display technology," *Computer*, vol. 38, no. 8, pp. 46–53, 2005.

[51] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. 1999 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR '99)*, vol. 2, pp. 23–25.

[52] A. Sullivan, "3-deep: new displays render images you can almost reach out and touch," *IEEE Spectr.*, vol. 42, no. 4, pp. 30–35, Apr. 2005.

[53] M. Tidwell, R. Johnston, D. Melville, and T. Furness, "The virtual retinal display—A retinal scanning imaging system," in *Proc. Virtual Reality World*, 1995, pp. 325–333.

[54] Y. Visell, A. Law, J. Ip, R. Rajalingham, S. Smith, and J. R. Cooperstock, "Interaction capture in immersive virtual environments via an intelligent floor surface," in *Proc. IEEE Virtual Reality*, Mar. 2010, pp. 313–314.

[55] Y. Visell, A. Law, and J. R. Cooperstock, "Touch is everywhere: Floor surfaces as ambient haptic interfaces," *IEEE Trans. Haptics*, July–Sept. 2009, pp. 148–159.

[56] H. L. Wang, K. Sengupta, P. Kumar, and R. Sharma, "Occlusion handling in augmented reality using background-foreground segmentation and projective geometry," *Presence: Teleoper. Virtual Environ.*, vol. 14, no. 3, pp. 264–277, 2005.

[57] H. Windsor, *The Boy Mechanic: Volume 1: 700 Things For Boys To Do.* Chicago, Illinois: Popular Mechanics, 1916.

[58] W. Woszczyk, J. R. Cooperstock, J. Roston, and W. Martens, "Shake, rattle, and roll: Getting immersed in multisensory, interactive music via broadband networks," *Audio Eng. Soc.*, vol. 53, no. 4, pp. 336–344, Apr. 2005.

[59] M. Wozniewski, Z. Settel, and J. R. Cooperstock, "A framework for immersive spatial audio performance," in *Proc. New Interfaces for Musical Expression (NIME)*, 2006, pp. 144–149.

[60] M. Wozniewski, Z. Settel, and J. R. Cooperstock, "User-specific audio rendering and steerable sound for distributed virtual environments," in *Proc. Int. Conf. Auditory Display*, June 2007, pp. 98–101.

[61] Q. Wu, P. Boulanger, and W. F. Bischof, "Bi-layer video segmentation with foreground and background infrared illumination," in *Proc. 16th ACM Int. Conf. Multimedia (MM '08)*. New York: ACM, 2008, pp. 1025–1026.

[62] L. Zhang and W. Tam, "Stereoscopic image generation based on depth images for 3d tv," *IEEE Trans. Broadcast.*, vol. 51, no. 2, pp. 191–199, June 2005.

[63] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *IEEE. Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, 2008, pp. 1–8. [SP]